ARC Centre of Excellence in Population Ageing Research

Working Paper 2020/25

Multi-State Health Transition Modeling Using Neural Networks

Qiqi Wang, Katja Hanewald and Xiaojun Wang

Multi-State Health Transition Modeling Using Neural Networks

Qiqi Wang[1,*], Katja Hanewald[2], and Xiaojun Wang[1]

July 2, 2021

Abstract

This article proposes a new model that combines a neural network with a generalized linear model (GLM) to estimate and predict health transition intensities. We introduce neural networks to health transition modeling to incorporate socioeconomic and lifestyle factors and to allow for linear and nonlinear relationships between these variables. We use transfer learning to link the models for different health transitions and improve the model estimation for health transitions with limited data. We apply the model to individual-level data from the Chinese Longitudinal Healthy Longevity Survey from 1998–2018. The results show that our model performs better in estimation and prediction than standalone GLM and neural network models. We provide new estimates of the life expectancies for a range of population subgroups. We also describe a wide range of possible applications for further health-related research, including risk prediction using health claim data and mortality prediction based on individual-level mortality data.

JEL classification: C13; C53; G22; J11

Keywords: Neural networks, Transfer learning, Multi-state health transitions, GLM

Acknowledgments: The authors acknowledge financial support from the Australian Research Council Centre of Excellence in Population Ageing Research (CEPAR). They are also grateful for comments received from Bernard Wong, Andrés Villegas, Timothy Neal, Weikang Xu, and Chen Yan.

* Contact: Qiqi Wang, email: wangqiqi_ruc@outlook.com.
[1] School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District Beijing, 100872, P.R. China.
[2] School of Risk & Actuarial Studies and Australian Research Council Centre of Excellence in Population Ageing Research (CEPAR), UNSW Sydney, Sydney, NSW 2052, Australia.

1. Introduction

This article develops a new model that combines a neural network with a generalized linear model (GLM) to estimate and predict age-specific health transition. The new model allows for age effects, time trends, socioeconomic factors, and lifestyle behaviors to impact the transitions between different health states in a Markov framework. The model detects and incorporates linear and nonlinear relationships among the variables autonomously without the need to specify the functional form of these relationships. We introduce transfer learning to link the models for different health transitions. Our new model has broad applications in insurance, actuarial and health research. We describe several possible applications, including risk prediction using health claim data and mortality prediction based on individual-level mortality data.

We develop our model using the health trajectories of older individuals as an application. In our model, healthy older individuals can develop long-term care (LTC) needs or die. We focus on older individuals because, around the world, people are living longer. With increasing age and longevity, individuals face a higher risk of chronic disease and age-related disability from chronic diseases, cognitive impairment, and functional limitations. As the number of older persons grows along with their longevity, the need for LTC will significantly increase in both developed and developing countries (United Nations, 2016). Therefore, strategies for how to provide and fund these growing LTC needs are needed. Our model can inform the development of such strategies by predicting the chance of individuals becoming disabled and needing LTC.

Our article makes several methodological contributions to the literature. First, we introduce socioeconomic and lifestyle factors in the modeling of health transitions. Many existing studies on multi-state health transition models use GLMs with a limited number of factors such as age, time, and LTC duration (see, e.g., Fong et al., 2015; Li et al., 2017; Fuino and Wagner, 2018; Hanewald et al., 2019). However, demographic and medical research has shown that socioeconomic and lifestyle factors also impact health states and health transitions. For example, there are relationships between health transitions and: marital status (Goldman et al., 1995; Robards et al., 2012); residency (Eberhardt and Pamuk, 2004; Ikeda et al., 2019); smoking behavior (Nusselder et al., 2000; Husemoen et al., 2004); and alcohol consumption (Oslin, 2000; Engchuan et al., 2019). Surprisingly, however, few studies include these variables in health transition models.

Second, we introduce neural networks to health transition modeling to incorporate

socioeconomic and lifestyle factors and to allow for linear and nonlinear relationships between these variables. Most of the current literature on multi-state health transition models relies on GLMs. The GLM framework provides a flexible approach to actuarial graduation techniques (Renshaw, 1991), which was first applied in health transition models by Renshaw and Haberman (1995). Recent papers that use GLMs to model health transitions include Fong et al. (2015), Shao et al. (2017), and Hanewald et al. (2019). We build on a growing literature that combines GLMs with neural networks. Wüthrich and Merz (2019) introduce the combined actuarial neural network (CANN) approach, which blends neural networks and classical GLMs. Schelldorfer and Wüthrich (2019) use the CANN approach to calculate claims frequency based on French motor third-party liability insurance data. They find that CANN enhances GLM by allowing for interactions between the variables. Gabrielli et al. (2020) also show how a GLM can be embedded into a neural network architecture to improve the over-dispersed Poisson model for general insurance claims reserving. However, we are the first to combine GLMs and neural networks to model health transitions. In doing so, we modify the CANN approach for health-related data. We show how expert opinion—for example, that age and time are key variables impacting health transitions—can be incorporated in a combined GLM–neural network model.

Third, we are the first to introduce transfer learning to link the models for different health transitions. Transfer learning aims to solve a problem by utilizing the knowledge learned from another problem. It is an important approach in machine learning to deal with small data size and enhance model performance by studying similar tasks. Our application involves modeling the health transitions between different health states. Previous studies on multi-state health transition modeling have modeled the different health transitions separately (e.g., Fong et al., 2015; Shao et al., 2017; Fuino and Wagner, 2018; Hanewald et al., 2019). However, as the health states are linked to each other, the morbidity and mortality transitions are also potentially linked (e.g., Alter and Riley, 1989; Johansson, 1991). Therefore, we introduce transfer learning to link the models for different health transitions. At the same time, transfer learning makes full use of the data by combining the data for different health transitions, which is helpful where data are limited.

We illustrate our model using data from the Chinese Longitudinal Healthy Longevity Survey (CLHLS), which has one of the largest samples of the oldest old in the world. Our sample consists of 69,063 observations for individuals aged 65–105 in eight survey waves over the period 1998–2018. The CLHLS collects intensive individual interview data, including health,

disability, demographic, family, socioeconomic, and behavioral risk factors for mortality and healthy longevity. The mortality and morbidity data from the CLHLS are high quality (Zheng, 2020) and have been used to model health transitions using GLMs (e.g., Hanewald et al., 2019).

We choose to illustrate our model based on data from China because China is experiencing very rapid population aging and has implemented a range of policies in response. In 2012, China piloted its first public long-term care insurance (LTCI) program in the city of Qingdao, which provided professional LTC and geriatric services for those with substantial LTC needs. Since 2016, the pilot program has been extended to over 40 cities in China. China is also promoting the establishment of a multi-pillar LTCI system and is encouraging the development of private LTCI products. Our new model provides a novel approach for predicting more accurate health transition estimates, which can inform the development of public and private LTCI.

The empirical results show that our new model performs better than a standalone GLM model or a standalone neural network. The new combined model reduces in-sample and out-of-sample losses and yields more accurate estimation and prediction results than standalone GLM models or standalone neural network models. Transfer learning improves the model by linking the separate models for the health transitions, which confirms that the transitions should be modeled together. The results also confirm that socioeconomic factors and lifestyle behaviors are significant and should be taken into account when estimating and predicting health transitions. Based on the estimated health transition intensities for certain groups of people with specific socioeconomic and lifestyle characteristics, we provide new estimates and predictions of the life expectancies and healthy life expectancies for different population subgroups.

In summary, this article shows that new models that combine traditional insurance techniques and neural networks can be used to model health transitions. We find that a combined GLM-neural network model that includes several socioeconomic and lifestyle factors, takes expert opinion into account and uses transfer learning to link the models for different health transitions outperforms other model variants. We discuss a wide range of possible applications of our modeling approach in insurance, actuarial, and health research in the final section of this paper.

The remainder of this article is structured as follows. In Section 2, we describe the methodology. In Section 3, we describe the CLHLS data and model settings. Section 4 presents the model results and the life expectancy estimates. Section 5 reports the sensitivity analysis. We conclude

the article in Section 6.

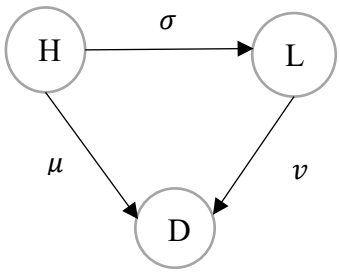
2. Methodology

This article introduces a new model to estimate and predict health transition intensities. In the following, we describe the components of this new model: the Markov process, the neural network component, how we combine the neural network with a GLM, how we incorporate expert opinion and transfer learning. Section 2.7 summarizes our proposed new model.

2.1. Markov Process

We assume that an individual's health transitions can be modeled as a multi-state Markov process, where the conditional probability distribution of future states of the process is independent of the process history and only depends on the state presently occupied. We consider a three-state Markov process with states "H" (healthy/no LTC needs), "L" (LTC needs), and "D" (dead). The process is shown in Figure 1. The three transition intensities are σ: the intensity of becoming LTC dependent; μ: the mortality intensity for a healthy person; and v: the mortality intensity for a person with LTC needs.[3]

Figure 1. Three-State Markov Process.



We assume that the Markov process is time-continuous and inhomogeneous with transition probabilities and transition intensities respectively denoted by

$$P_{ij}(\boldsymbol{x}_g, t, u) = P_r\{S(\boldsymbol{x}_g, u) = j | S(\boldsymbol{x}_g, t) = i\}, \;\; 0 \le t \le u, i, j \in \{H, L, D\} \tag{1}$$

$$TI_{ij}(\boldsymbol{x}_g, t) = \lim_{u \to t^+} P_{ij}(\boldsymbol{x}_g, t, u)/(u - t)\,, t \ge 0, i \ne j\,, \tag{2}$$

[3] Hanewald et al. (2019) use a similar three-state Markov process to model health transitions using GLMs based on CLHLS data. In the CLHLS data, there are very few recovery transitions from the functionally disabled to the nondisabled state. Therefore, we do not consider such recoveries in this study.

where $\boldsymbol{x}_g$ is a vector containing features (or variables)[4] with:

$$\boldsymbol{x}_g = (\text{Age}_g, \text{Gender}_g, \text{Residency}_g, \text{Marital status}_g, \text{Smoking}_g, \text{Drinking}_g)'.$$

The index g denotes a group of people with a certain value for each variable. t is time with $h \geq 0$. $S(\boldsymbol{x}_g, t)$ denotes the stochastic health status of an individual with certain features and time t, and $i, j \in \{H, L, D\}$. $P_{ij}(\boldsymbol{x}_g, t, u)$ denotes the transition probability from state i with certain features $\boldsymbol{x}_g$ and time t to state j at time $t + h$. $TI_{ij}(\boldsymbol{x}_g, t)$ is the instantaneous transition intensity. TI_{ij} is age-dependent and is assumed to be integrable on compact intervals.

We would like to briefly discuss some of the model assumptions:

The number of health states: We model three health states (healthy/no LTC needs, LTC needs, and dead). Our model can be easily extended to more health states for other applications, provided that sufficient data exist. A version of the model with two health states (alive, dead) can be used to model mortality rates.

Markov process: We model an individual's health transitions as a multi-state Markov process, where the transition probability only depends on the individual's current health state. We acknowledge that, in practice, past health states can impact LTC disability and mortality rates so that the health transitions could be modeled as semi-Markov processes. We adopt the Markov assumption for three reasons: (i) to reduce complexity and obtain an easily manageable model (Christiansen, 2012; Biessy, 2017); (ii) to facilitate comparison with previous studies (e.g., Fong et al., 2015; Shao et al., 2017; Hanewald et al., 2019); and (iii) because there is no information on LTC disability duration reported in the data set we used and the time intervals between survey waves are too long to estimate disability duration. Our model could be easily applied under a semi-Markov assumption when data on LTC disability duration are available.

Time-inhomogeneous Markov process: We allow the health transition intensities to be dependent on time to account for time trends.

Continuous-time Markov model: The panel dataset we use contains the current health status of the survey participants at the time of each follow-up survey. The time intervals between surveys vary between two to four years. We approximate the underlying continuous health process by a continuous-time Markov model. This model assumes that an individual will stay in one health state for a period of time and then immediately jump to another state. The continuous-time

[4] We use the terms "variables" and "features" interchangeably.

Markov process allows for equal and unequal observation intervals, which is suitable for longitudinal surveys with different observation intervals. As a result, the continuous-time Markov model is widely used in the actuarial literature using follow-up survey data (see, e.g., Jones and Grunwald, 2006; Fong et al., 2015; Hanewald et al., 2019).

2.2. Generalized Linear Models

The standard approach for modeling the health transitions in Figure 1 is to estimate separate standalone GLMs for each transition intensity (e.g., Renshaw and Haberman, 1995; Fong et al., 2015; Shao et al., 2017; Hanewald et al., 2019).

Let n_g denote the number of transitions in group g and and e_g denote the corresponding central exposure to risk in each interval. In common with Renshaw and Haberman (1995), we assume that a transition intensity is constant within one-year age intervals, and thus n_g follows a Poisson distribution

$$n_g \sim Poi(\eta(\boldsymbol{\gamma_g}) e_g), \tag{3}$$

with $\boldsymbol{\gamma_g} = (\mathrm{t}_g, \mathrm{Age}_g, \mathrm{Gender}_g, \mathrm{Residency}_g, \mathrm{Marital\ status}_g, \mathrm{Smoking}_g, \mathrm{Drinking}_g)'$ and $\eta(\boldsymbol{\gamma_g})$ represents the transition intensity.

The GLM is calibrated to the data based on independent Poisson response variables n_g, where

$$E(n_g | e_g, \eta(\gamma_g)) = e_g \eta(\gamma_g) = m_g, \tag{4}$$

$$Var(n_g | e_g, \sigma x) = \varphi m_g. \tag{5}$$

Thus, the GLM takes the form:

$$log\, \eta(\boldsymbol{\gamma}) = \beta_0 + \textstyle\sum_{l=1}^{q_0} \beta_l \gamma_l \overset{\text{def}}{=} \langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle \tag{6}$$

for the parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q_0})' \in \mathbb{R}^{q0+1}$, $\boldsymbol{\beta}$ stands for unknown coefficients that need to be estimated, $\boldsymbol{\gamma}$ represents the variables we have in the model, and q_0 is the number of chosen variables.

The GLMs are fitted to the data using maximum likelihood estimation to estimate the regression coefficients $\boldsymbol{\beta}$ and their associated asymptotic standard errors.

As noted in the introduction, many existing studies on multi-state health transition models use GLMs with a limited number of factors such as age, time, and LTC duration (see, e.g., Fong et al., 2015; Li et al., 2017; Fuino and Wagner, 2018; Hanewald et al., 2019). We include a GLM

(GLM0) that only includes age and time with an interaction between age and time (as in Hanewald et al., 2019) in the model comparison:[5]

$$\eta_{age,time} = \beta_0 + \beta_1 age + \beta_2 time + \beta_3 age * time. \quad (7)$$

We also include a GLM with additional covariates in the model comparison. In doing so, we extend the current literature in this area.

2.3. Neural Networks

In this section, we describe the neural network component of our model. Neural networks are machine learning algorithms that are inspired by the biological neural networks in the human brain. They are versatile and powerful machine learning techniques and are widely used in pattern recognition, recommendation systems (e.g., to recommend music clips in different apps), and medical diagnoses. There is also a growing research interest in neural networks in insurance and actuarial studies (e.g., Brockett et al., 2006; Cheng et al., 2020; Kiermayer and Weiß, 2020).

Neural network models consist of nodes that connect to other nodes that have associated weights. The nodes are aggregated into layers. There is one input layer of nodes, one or more hidden layers, and an output layer. Neural networks with more than one hidden layer are called deep neural networks.

At a node, the inputs from previous nodes are multiplied by the weights (which are estimated from the data) and summed up. The sum is passed on to an activation function, which further modifies the output before passing it on to the next node. The type of activation function is a model setting, and there are many different types of activation functions. The information passed on by the activation function depends on some thresholds being assumed by the function.

The information from the nodes in the input layer is processed through all nodes in the next layer, with different weights at each node. For neural networks with several hidden layers, this process is repeated at each layer. Thus, the hidden layers perform nonlinear transformations of the inputs entered into the network.

[5] We note that when we only consider age and time, we calculate the transition intensities using different counts of transition and exposure. Thus, we only group the individual data based on different ages and time. However, when we take other variables into account, the counts need to be subdivided, and exposures are calculated according to the given values of all these variables.

In this paper, we focus on regression neural networks, which have a regression function as an activation function for the output layer. Another type of neural network, classification neural networks, has different activation functions for the output layer.

We start by defining our health transition model in a neural network framework:

Our Poisson regression framework in the form of neural networks is given by

$$log\,\eta(\boldsymbol{\gamma}) = \left\langle \boldsymbol{w}^{(K+1)}, (\boldsymbol{z}^{(K)} \circ \cdots \circ \boldsymbol{z}^{(1)})(\boldsymbol{\gamma}) \right\rangle, \tag{8}$$

with $\boldsymbol{q_k}$ hidden neurons in the k-th hidden layer given by

$$z_j^{(k)}\ (\boldsymbol{z}) = f(\boldsymbol{w}_{j,0}^{(k)} + \textstyle\sum_{l=1}^{\boldsymbol{q_{k-1}}} \boldsymbol{w}_{j,l}^{(k)}\, z_l), \tag{9}$$

where $\boldsymbol{\gamma}$ represents the variables we have in the model, $\boldsymbol{w}^{(k)}$ represents the weights, and f is a (nonlinear) activation function.

We can apply a neural network to model the health transition intensities because the *universal approximation theorem* ensures that a neural network with a single layer, a finite number of nodes, and an activation function can approximate any arbitrary complex and continuous relationship among the input variables (Goodfellow et al., 2013).

We include two standalone neural network models in the model comparison: NN0, which only includes age and time, and NN, which includes age, time, and additional covariates. For both NN0 and NN, we use a deep neural network structure to detect linear and nonlinear relationships between the variables.

2.4. Basic Combined Model

To develop our proposed new model, we combine a neural network with a GLM and include important variables other than age and time to enhance the model's estimation performance.

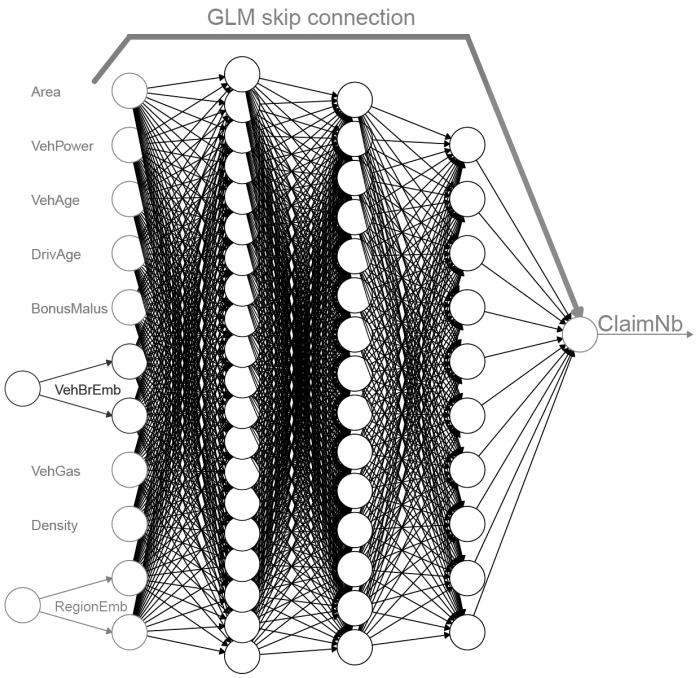
As noted previously, Wüthrich and Merz (2019) and Schelldorfer and Wüthrich (2019) use the CANN approach, which nests a GLM into a neural network architecture. The CANN model can be defined as follows:

$$log\,\eta(\boldsymbol{\gamma}) = \langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle + \left\langle \boldsymbol{w}^{(K+1)}, (\boldsymbol{z}^{(K)} \circ \cdots \circ \boldsymbol{z}^{(1)})(\boldsymbol{\gamma}) \right\rangle. \tag{10}$$

Equation (10) combines the regression function from the GLM in equation (6) and the regression function from a neural network given in equation (8). Figure 2, taken from

Schelldorfer and Wüthrich (2019), shows the original CANN approach.

Figure 2. Original CANN Approach.



Note: The line represents the GLM in the skip connection added to a neural network.

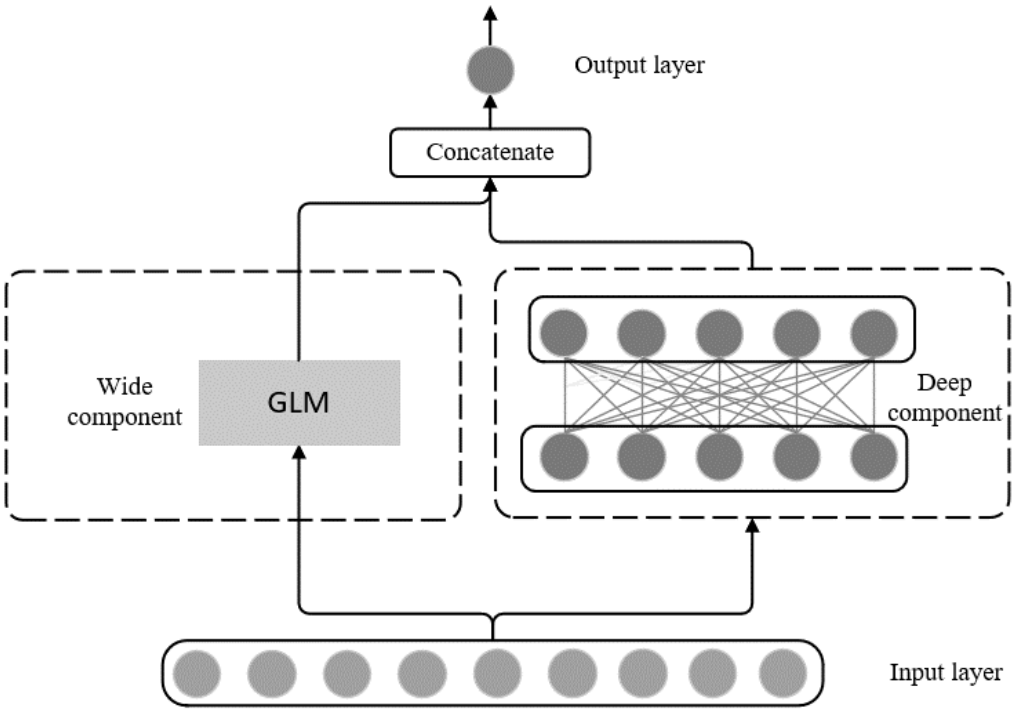
Source: Schelldorfer and Wüthrich (2019, p. 14).

Figure 3 shows the combined model (CM) in our paper. We modify the CANN approach for health-related data. We combine a GLM with a neural network to construct a wide and deep network architecture (Cheng et al., 2016). The 'architecture' of our combined model differs from Schelldorfer and Wüthrich's (2019) CANN approach shown in Figure 2. We note that CM is a non-trivial adaptation of the CANN framework to health-related data.

In our model, the deep component is a neural network with several hidden layers that can find unknown variable combinations (i.e., the linear and nonlinear connections) among the socioeconomic and lifestyle variables. The wide component is a GLM, as described in Section 2.2. The estimated parameters of the GLM pass on to the last hidden layers as the wide component. The GLM part of the model does not include any interactions between the variables because the neural network finds linear and nonlinear interactions. The outputs of the deep component and the wide component concatenate (are combined) using a logistic function, and

the resulting information is passed on to the activation function in the output layer.

Figure 3. Wide and Deep Structure in the Combined Model.

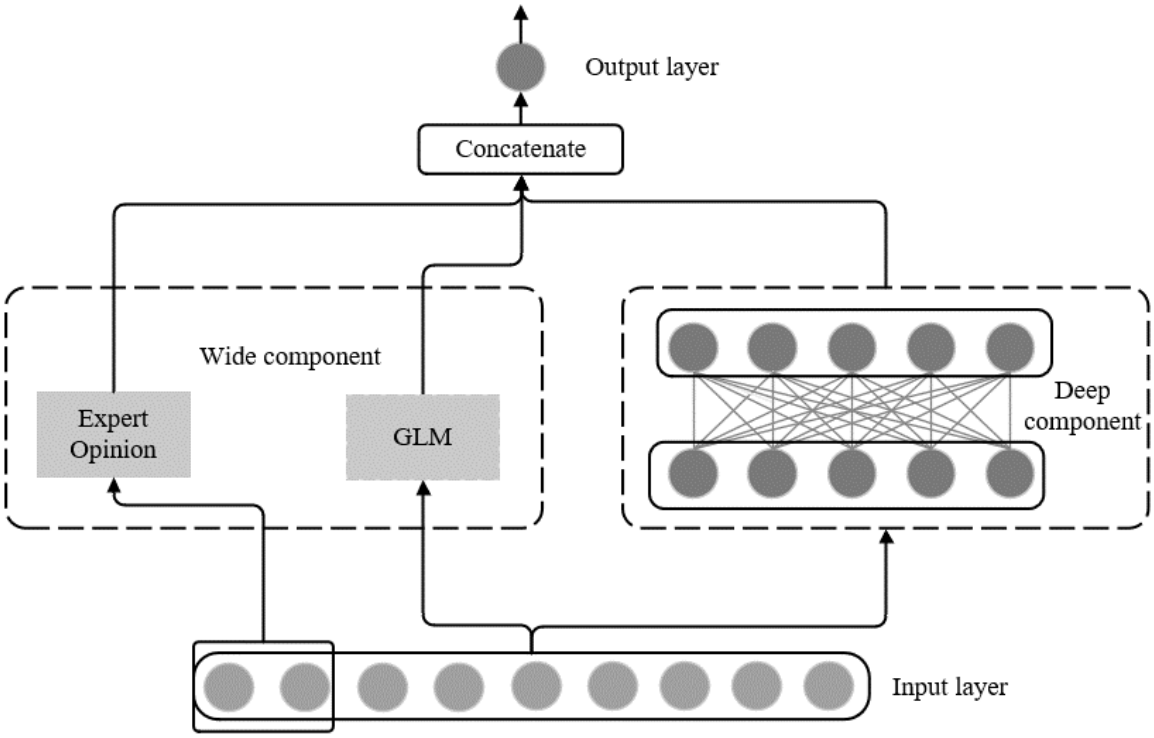


2.5. Expert Opinion

We can incorporate expert opinion into a wide and deep network structure (Alashkar et al., 2017). To emphasize the importance of specific variables, we can pass specific variables to an activation function and then add it directly to the last hidden layer to reflect the relationship between the variable and the outputs. In previous studies, age patterns and time trends have typically been included to model health transitions (Renshaw and Haberman, 1995; Fong et al., 2015; Shao et al., 2017; Hanewald et al., 2019). Building on this literature, we add age and time into the last layer of our model and jointly train[6] the wide component and deep component to further improve our model. This structure is shown in Figure 4. We call this model variant 'combined model with expert opinion (CME)'.

[6] Training is the process of finding the appropriate weights in neural networks.

Figure 4. Combined Model with Expert Opinion.



2.6. Transfer Learning

A key innovation of our model is that it features transfer learning between different health transitions. Transfer learning is a machine learning method that improves learning in new tasks by transferring knowledge from related tasks that have been learned (Ye and Dai, 2018; Zhuang et al., 2020). [7] In other words, transfer learning utilizes the knowledge from one task to solve related tasks. It reuses the trained model as the starting point to build a new model for a related task. We refer to Tan et al. (2018) for a mathematical definition of transfer learning and a description of different transfer learning techniques.

Most previous research on transfer learning is related to object detection, image classification, and various medical applications. Most of the research has been published in computer science journals. A few recent studies have used transfer learning in finance-related applications, for

[7] Transfer learning differs from credibility theory, which is an important actuarial tool for estimating pure premiums, frequencies, and severities for individual risks or classes of risk. Credibility theory combines multiple estimates of a future event to achieve a more accurate estimate. Transfer learning uses pre-trained models as the starting point for the new but similar task. Transfer learning relaxes the assumption that training data and future data must be in the same feature space and have the same distribution. Transfer learning can use different types of knowledge (features, weights, etc.) from previously trained models to train newer models.

example, to improve financial trading decisions (Jeong and Kim, 2019) and in text analysis of financial information (Kraus and Feuerriegel, 2017; Tang et al., 2019).

Transfer learning is mainly used to enhance the performance of neural networks, but it can also be applied to other machine learning methods, for example, in boosting algorithms (Dai et al., 2007) and deep forests (Utkin and Ryabinin, 2017).

We introduce transfer learning into the model for two reasons. First, we want to model different health transitions together instead of modeling them separately. While the health states in our model link directly to each other, as shown in Figure 1, there can also be connections between the transitions (see, e.g., Alter and Riley, 1989; Johansson, 1991). Transfer learning can transfer knowledge from one health transition to other transitions to improve model performance. Second, in the successful CANN application examples (Schelldorfer and Wüthrich, 2019; Gabrielli, 2020), the data sets are relatively large, and a model is built for one sample. However, health-related data are often limited. Transfer learning can be used to develop models for limited data by learning from related models based on related (larger) datasets. In health transition studies, we often model several health transitions related to different health states. When these health transitions are modeled separately, the data for establishing each model can be relatively limited. Transfer learning allows us to use all available data (for all health transitions) to model specific health transitions.

There are several transfer learning methods (see, e.g., Zhuang et al., 2020), which differ in the kind of knowledge that is transferred from one task to a related task. We choose the parameter transfer method, in which the related models share layer parameters (Kim and Youn, 2019). The parameter transfer method encodes transferred knowledge into these parameters. Specifically, the shared parameter-based transfer method attempts to find the common parameters between related tasks and transfer these parameters from the source task to the target so that knowledge transfer can be achieved through further processing neural networks (Kumagai, 2016). We use the method to model the relationships between the three transition intensities in our Markov model (see Figure 1). With this approach, the potential links between different health transitions lie in the parameter transfer, and transfer learning can improve the performance of our new model by learning useful knowledge from other transitions during model training.[8]

[8] Learning in transfer learning is the passing on of knowledge from a trained model to a related task.

We estimate one model for the health transition with the largest number of observed transitions, which is the mortality intensity for a healthy person μ, and fine-tune the model.[9] Then we transfer the weights of the trained model as the initialization for the models for the other two health transitions (i.e., the intensity of becoming LTC dependent σ and the mortality intensity for a person with LTC needs ν). We use transfer learning in two model variants in the model comparison: in the combined model with transfer learning (CMT) and the combined expert opinion neural network model with transfer learning (CMET).

2.7. Proposed New Model

Our final model (CMET) includes all the elements described in the previous subsections: we combine a GLM with a neural network, add covariates to both the GLM and the neural network parts of the model, include expert opinion, and use transfer learning to model different health transitions together and to improve the model performance. Our approach does not establish causality but finds the relationships between outputs (health transition intensities) and inputs (socioeconomic and lifestyle factors). In section 4, we will compare our proposed new model with different GLMs, neural networks, and combined models without transfer learning.

2.8. Estimation and Loss Functions

To estimate the models mentioned in the previous subsections, we randomly split our data into a training data set D and a test data set T. The GLM models are fitted with maximum likelihood estimation to the training data set D. The log-likelihood ℓ is

$$\ell = \sum_{g=1}^{N_D} n_g \, log\left(\eta(\boldsymbol{\gamma_g})\right) - e_i \eta(\boldsymbol{\gamma_g}) \quad , \tag{11}$$

where N_D is the sample size in the training data set D.

We assess the quality of the models based on the mean square error (MSE), which is a commonly used loss function to train models and measure prediction performance in deep learning (see, e.g., Yin and Htay, 2020; Bu et al., 2021). We compare models based on their in-sample MSE given by formula (12) using the training data set D and the out-of-sample MSE given by formula (13) using test data set T. The in-sample loss assesses the fit and the out-of-

[9] The objective of fine-tuning is to adjust the weights of the trained model from the final phase to improve the prediction outcome (Lin et al., 2018).

sample loss assesses the prediction performance.

$$L(D,\hat{\eta}) = \frac{1}{N_D}\sum_{g=1}^{N_D}\left(log\left(\frac{n_g}{e_g}\right) - log\left(\hat{\eta}(\boldsymbol{\gamma_g})\right)\right)^2. \quad (12)$$

$$L(T,\hat{\eta}) = \frac{1}{N_T}\sum_{g=1}^{N_T}\left(log\left(\frac{n_g}{e_g}\right) - log\left(\hat{\eta}(\boldsymbol{\gamma_g})\right)\right)^2. \quad (13)$$

3. Data and Model Settings

3.1. Data

We illustrate our model using longitudinal data from the CLHLS, a follow-up survey organized by an international team of researchers with funding and institutional support from a range of sources, including the U.S. National Institute on Aging; Peking University; and the United Nations Fund for Population Activities (UNFPA). Data are available from eight survey waves over the years 1998–2018. The data set can be download for free after registration (see Center for Healthy Aging and Development Studies, 2020).

The CLHLS focuses on the oldest-old aged 80 and older from 22 provinces in mainland China. Since 2002, the CLHLS has included a comparison group of younger elders aged 65–79. The CLHLS uses face-to-face interviews based on internationally compatible questionnaires to collect detailed information about health, socioeconomic characteristics, family, lifestyle, and other demographic variables. At each wave, the survivors are re-interviewed, and deceased interviewees are replaced with new participants. Data on mortality and health status before dying are collected in interviews with a close family member of the deceased. For more detailed information about the survey design, please refer to Zheng (2020).

The CLHLS has one of the largest samples of the oldest-old in the world. Even at higher ages, the sample size of the CLHLS is large. We use data for the 65–105 age group in one-year age groups. We discard 840 individuals with missing independent and dependent variable information, leaving a baseline sample of 69,063 observations in eight survey waves. In some waves, a small amount of data is collected in the next calendar year. We consider the exact date of the interview when calculating the risk exposure.

We use information about ADL (activities of daily living) limitations to classify individuals into LTC states. In all waves of the CLHLS, six ADL items were consistently evaluated: bathing, dressing, eating, using the toilet, continence, and transferring in and out of bed.

Individuals rated their ability to perform these activities on a three-point scale (1 = no help required, 2 = partial assistance required, 3 = full assistance required). We classify individuals as being able to perform an ADL only when they need no assistance. If they encounter difficulties with two or more ADL, we define them as LTC dependent, which is consistent with previous research by Fong et al. (2015) based on U.S. data, and Hanewald et al. (2019) based on data from the CLHLS. This definition of LTC disability is also consistent with the trigger conditions for the payment of benefits for many existing LTCI policies in the U.S. market and some LTCI pilot programs in China.

Table 1 summarizes the variables (or factors) included in our model based on previous research. The table also summarizes how we pre-processed the data following previous studies. We selected variables that (i) were available in the CLHLS data, (ii) are known to be related to the health transition based on previous studies (see, e.g., Goldman et al., 1995; Robards et al., 2012; Engchuan et al., 2019), and (iii) can be easily verified by insurance companies. The model can be easily extended to include additional variables if larger and more detailed data are available.

Table 1. Variables.

Variable	Reference	Type of variable	Details
Age	Fong et al. (2015)	Continuous	One-year age groups 65–105
Time	Li et al. (2017)	Continuous	Seven starting time points: 1998, 2000, 2002, 2005, 2008, 2011, 2014
Gender	Hanewald et al. (2019)	Binary	Rural vs. urban residency
Residency	Hanewald et al. (2019)	Binary	Rural vs. urban residency
Marital status	Gu and Yi (2004)	Binary	Currently married and living with spouse vs. not
Smoke	Gu and Feng (2015)	Binary	Currently smoking vs. not
Drink	Gu and Feng (2015)	Binary	Currently drinking vs. not

Note: All references above are for processing the variables.

We use data from all eight CLHLS survey waves (1998, 2000, 2002, 2005, 2008, 2011, 2015, 2018). To make full use of the available information, we use an unbalanced panel design that includes all individuals with at least two consecutive observations. As a result, each individual can have up to seven health transitions. And described in Table 1, we use time as a continuous variable, which is consistent with previous research on health transitions using CLHLS data (Hanewald et al., 2019). For the model estimation, we define the year 1998 as $t = 0$ and set the data points in the model to $t = (1, 3, 5.5, 8.5, 11.5, 14.5, 18)$ to reflect the fact that the transition intensities refer to the middle of the time intervals between survey waves and to account for the different interval lengths between survey waves.

Table A.1 in the appendix reports the attrition rates for each wave compared to the previous wave. The attrition rates vary between 8% and 21%. The attrition rates in the CLHLS are similar to that in surveys conducted in Western countries (Gu, 2008). The CLHLS does not provide weights that would allow us to correct attrition. We acknowledge that as a result, we are probably underestimating the three transition intensities σ, μ and ν because individuals who could not be followed up might have died without being recorded, or they may have been institutionalized in a hospital or nursing home. We recommend that future research should check and correct for attrition if data are available.

We calculate crude transition intensities as the number of health transitions divided by the corresponding central exposure to risk for a given time interval and given values of other variables. Table 2 reports that there are 31,660 transitions in the dataset, of which 15% are transitions into LTC needs, 59% are the deaths of healthy individuals, and 26% are the deaths of individuals with LTC needs. The total number of exposure years is 156,875. We also compare the transition counts and exposures of age 85 and age 105 in Table A.2 in the appendix. The comparison shows that there are fewer transition counts at higher ages, and therefore the crude transition intensities vary more at higher ages. Table A.3 in the appendix shows the crude intensities for each wave and each age group. The intensities increase by age and fluctuate over time.

Table 2. Transition Counts for Different Variables.

Variables		Transition counts			Exposure years	
		σ: H→L	μ: H→D	ν: L→D	H	L
Gender	Male	1,758	8,457	2,453	61,876	7,446
	Female	3,145	10,118	5,729	70,499	17,053
Marital status	With spouse	1,011	3,426	875	47,913	3,922
	Without spouse	3,892	15,149	7,307	84,463	20,578
Residency	Rural	2,540	11,103	4,431	75,867	12,917
	Urban	2,363	7,472	3,751	56,509	11,583
Smoke	Yes	730	3,385	638	27,781	2,466
	No	4,173	15,190	7,544	104,594	22,033
Drink	Yes	834	3,761	940	28,872	3,226
	No	4,069	14,814	7,242	103,504	21,273
Total		4,903	18,575	8,182	132,376	24,500

We show the correlations between the variables and the health transition intensities in Figure A.1 in the appendix. Unlike intensity v, intensities σ and μ have strong positive correlations with age. Our model can capture these linear and other nonlinear relationships between the variables.

3.2. Model Settings

We estimate the standalone neural networks NN0 and NN and the deep component in the models CM, CME and CMET with three hidden layers and 80 nodes for each layer. Section 5 reports the results of a sensitivity analysis regarding the choice of these hyperparameters. We use the scaled exponential linear unit (SELU) as the activation function (Klambauer et al., 2017). We use the SELU for all hidden layers because it ensures that our network self-normalizes.[10] We use linear activation in the last layer to solve the Poisson regression problem.

We use a stochastic gradient descent[11] method called adaptive moment estimation (Adam) (Kingma and Ba, 2014) to search for optimal weights. Adam is straightforward to implement and computationally efficient, and for this reason, it is a popular method in deep-learning applications. The weights are tuned by backpropagation.[12]

Deep neural networks have a large number of parameters that can cause overfitting. We use the dropout technique to avoid overfitting (Hinton et al., 2012). The idea of dropout is to randomly drop nodes from the neural network during training. In this article, we set the dropout probability at 10%.

We adopt batch normalization for faster and more stable training of the model[13]. Proposed by Ioffe and Szegedy (2015), batch normalization is a method used to make neural networks faster and more stable by normalizing the input layer by re-centering and re-scaling. Batch

[10] Self-normalization means that the output of each layer will preserve a mean of zero and a standard deviation of one during training, which solves the vanishing/exploding gradients problem. The exploding gradient problem is when the derivatives are large and the gradient increases exponentially through the neural network.
[11] Gradient descent is an optimization algorithm that minimizes the loss function by repeatedly moving along the steepest descent defined by the negative value of the gradient.
[12] Back propagation is a gradient descent optimization algorithm. Each weight is adjusted based on the loss back propagated from the output to the input.
[13] In the neural network training process, the entire dataset passes through the network multiple times to adjust the weights. To process the data into the network for training, the dataset is divided into a number of batches. One epoch is when the whole dataset passes through the network once, and the size of the batch is the number of samples in one batch.

normalization subtracts the batch mean and divides it by the batch standard deviation. It can also eliminate the need for dropout, so we only add dropout to the last hidden layer (Li et al., 2018).

We implement the neural network models using Google's 'TensorFlow' package interacted with the programming language Python and estimate the GLM models using the Python module 'statsmodels'. We are happy to share the code and provide guidance on how to implement the models.

4. Empirical Results

4.1. Model Comparison

In the following, we compare our proposed new model against several alternative models. Table 3 shows the average computing time and average losses for the following models (all variables are described in Table 1):

- **GLM0** is a GLM that only includes age and time with an interaction between age and time.
- **GLM** is a GLM that includes all the variables presented in Table 1.

Table A.4 and Table A.5 in the appendix report the parameter estimates for GLM0 and GLM .

- **NN0** is a neural network that only contains age and time as input variables.
- **NN** is a neural network that includes all the variables listed in Table 1.
- **CM** is the basic combined model described in Section 2.4 that combines NN and GLM.
- **CME** is the combined model with expert opinion incorporated: age and time are added to the last hidden layer, as described in Section 2.5.

The six models mentioned above model the three health transitions shown in Figure 1 separately. The final two models link the three health transitions by transfer learning as described in Section 2.6.

- **CMT** is a combined model with transfer learning.
- **CMET** is our proposed combined model with expert opinion and transfer learning.

Table 3 reports the in-sample and out-of-sample losses along with the computing times[14] for the different models for the three health transition intensities σ, μ, and ν. We first compare the

[14] The computing time was measured on a personal laptop Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz with 16GB RAM.

standalone GLM and neural network models. We note that GLM0 and NN0 have different counts and exposures than GLM and NN, so we cannot directly compare the losses. However, Appendix A.5 shows that all the variables in GLM are significant, which indicates that these variables should be included when modeling the health transitions. The standalone neural network NN performs much better than the GLM (and NN0 performs better than GLM0), as all in-sample and out-of-sample NN losses are much lower than for the GLM.

The basic combined model CM has smaller in-sample losses than the neural network NN, but the out-of-sample losses are larger for one transition intensity (ν) compared to NN. This suggests that a combined GLM-NN model can be suitable for some health transitions but can also be improved.

Adding expert opinion in model CME (in our case, adding age and time in the last layer) gives a better fit (lower in-sample loss) and better prediction (lower out-of-sample loss) than model CM and the standalone GLM and NN models for all three health transitions. Compared with model CM, the out-of-sample MSE losses for model CME are 1% lower for the intensity of becoming LTC disabled (σ), 3% lower for the mortality intensity of healthy individuals (μ), and 3% lower for the mortality intensity of individuals with LTC needs (ν). This shows that it can be important to take expert opinion into account and confirms that age and time are important when modeling health transitions.

Adding transfer learning to the combined model (in model CMT) also reduces both in-sample and out-of-sample losses compared to model CM, especially for the intensity of becoming LTC disabled (σ). The out-of-sample MSE losses for model CMT compared with model CM are about 17% lower for σ, the same for μ, and 4% lower for ν. This indicates that the health transitions should be modeled together.

Finally, the best performance in both in-sample and out-of-sample losses is achieved in our proposed combined model CMET, which incorporates both expert opinion in the wide component with age and time and transfer learning. Overall, our proposed model CMET decreases the out-of-sample loss compared with the basic CM model by about 25% for σ, 3% for μ, and 7% for ν.

Overall, our new model CMET has the best performance. Nesting classic actuarial models and expert opinion as the wide component into a neural network gives more accurate estimates by reducing in-sample and out-of-sample losses. Transfer learning connects the models for different transitions, which further improves the overall model. The results also confirm that

the three health transition intensities should be modeled together.

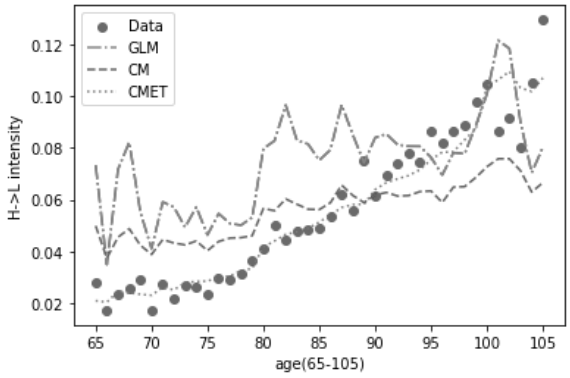
Table 3. MSE Loss Comparison and Computing Time for Different Models.

Models	In-sample loss ($\times 10^{-2}$)			Out-of-sample loss ($\times 10^{-2}$)			Computing time (s)		
	σ: H→L	μ: H→D	ν: L→D	σ: H→L	μ: H→D	ν: L→D	σ: H→L	μ: H→D	ν: L→D
GLM0	158.43	77.54	45.99	125.55	63.32	66.71	0.04	0.06	0.04
GLM	80.69	86.42	56.08	95.22	96.80	55.78	0.04	0.06	0.04
NN0	129.11	59.27	30.11	145.65	18.98	53.31	2.55	2.57	2.23
NN	43.03	31.20	52.86	49.29	34.40	53.37	10.27	14.96	9.34
CM	38.54	29.29	52.04	42.78	31.70	54.10	10.48	15.39	9.68
CME	37.32	28.50	51.01	42.34	30.82	52.67	11.68	17.06	10.79
CMT	31.27	29.29	50.22	35.35	31.70	52.02	10.54	15.39	9.65
CMET	31.08	28.50	49.70	32.25	30.82	50.47	11.73	17.06	10.82

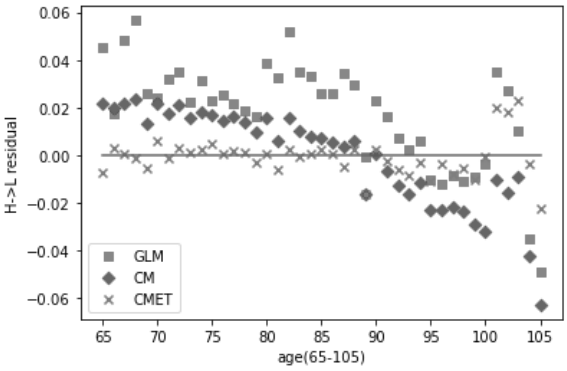
To illustrate how combined neural network methods and transfer learning improve the models, we compare the observed transition intensities against the fitted transition intensities based on the models GLM, CM, and CMET in Figure 6. To show the results only by age, we average the estimated intensities for all individuals within the same age group. To do so, we first calculate the estimated counts of health transitions and then divide the sum of the counts in one age group by the exposure in the same age group.[15] The intensity of becoming LTC disabled (σ) and the mortality intensity of healthy individuals (μ) increase by age. The estimated intensities fluctuate at higher ages due to limited data. CM and CMET smooth the curves by age, and the estimates of CMET are closer to the data.[16] The mortality intensity for individuals with LTC needs fluctuates before age 80 and then rises gently. Overall, CMET improves the GLM fitting, and transfer learning improves the models. The comparison shows that adding age and time in the last layer and introducing transfer learning improves the model fit of health transition intensities.

[15] We note that some of the fitted intensities shown in Figure 6 are more erratic than the data because they result from models that are fitted to all covariates. This is because the different models were estimated based on counts and exposures grouped by all covariates (not just age). Therefore, the counts and exposures in Figure 6 are only group by age which are different from the counts and exposures we used to estimate the health transitions.
[16] In practical applications, incidence rates and mortality rates can be smoothed by age and time.

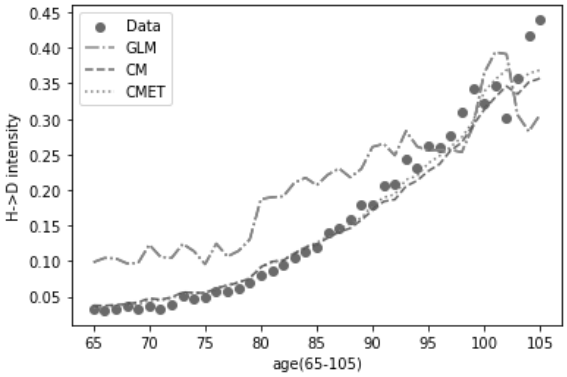
Figure 6. Intensities and Residuals by Age.



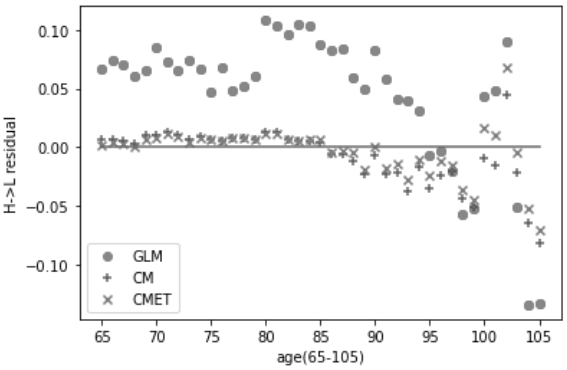
(a) σ: H→L intensity



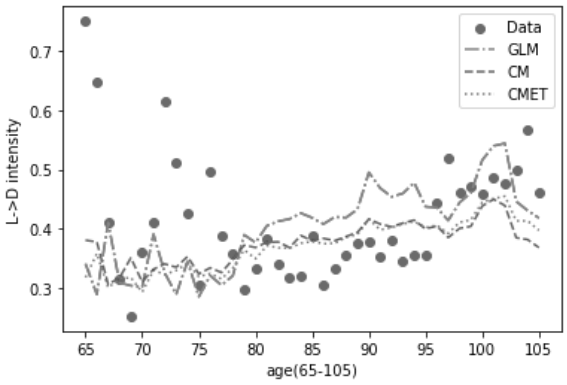
(b) σ: H→L residual



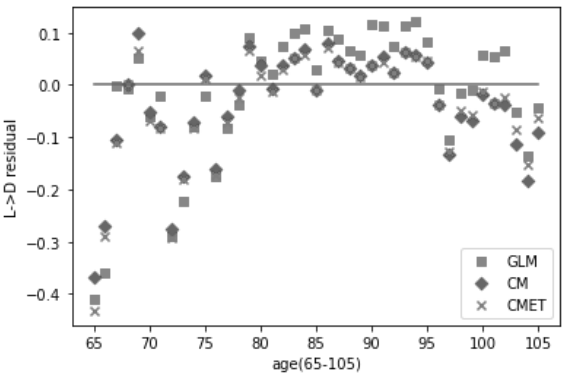
(c) μ: H→D intensity



(d) μ: H→D residual

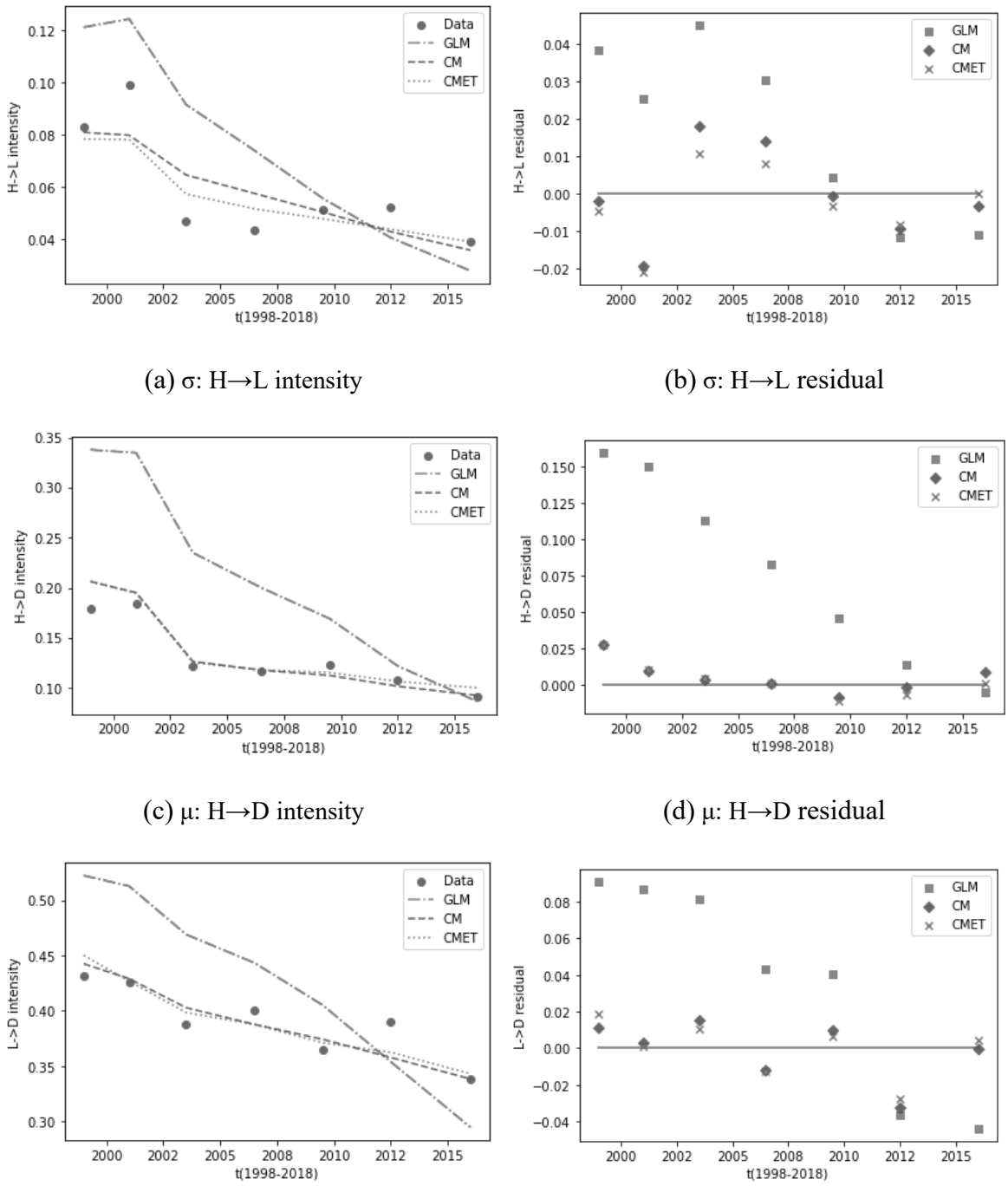


(e) ν: L→D intensity



(f) ν: L→D residual

Figure 7. Intensities and Residuals over Time.



(a) σ: H→L intensity

(b) σ: H→L residual

(c) μ: H→D intensity

(d) μ: H→D residual

(e) ν: L→D intensity

(f) ν: L→D residual

Figure 7 shows the estimated intensities and corresponding residuals over time. We note that the intensities decline over time, with some fluctuations between waves. There is a drop in some intensities from 2000 to 2002. A possible reason for this drop is that since 2002 the CLHLS has included a comparison group of younger older persons aged 65–79 who have lower transition intensities. Our model smooths the fluctuations and gives estimates and predictions

that are closer to the data compared with the GLM.

Figures A.2 and A.3 in the appendix show the estimates of CMET. Figure A.2 illustrates how variables other than age and time influence the health transition intensities. The three columns show the three different intensities. The first row compares the two genders. For most ages, males have higher transition intensities, reflecting a higher risk of becoming disabled or dying. However, there is a slight change after age 98 in the disability intensity, which shows that at a higher age, women are more likely to become disabled than men. The second row illustrates the effects of urban versus rural residency. Individuals in urban areas have higher mortality intensities but are less likely to be disabled than those in rural areas. This finding agrees with Wei and Wang (2017), who also employ CLHLS data. As for marital status, we find that healthy individuals who are presently living with a spouse have a lower risk of becoming disabled or dying (Zeng, 2013; Mu and Yang, 2016). However, once disabled, individuals living with a spouse have higher mortality intensities than those without a spouse. The next two rows show how lifestyle impacts health transitions. The intensities are much higher for individuals who smoke or drink alcohol compared to others, which confirms the adverse health effects of smoking and drinking.

Figure A.3 compares the average of the estimated intensities over time by gender and residency. The intensities decline over time, with different rates for males and females, and urban and rural residents. The results for other variables are available upon request. Overall, our model reflects the time trends in the health transitions and reasonable differences for different demographic groups.

4.2. Importance of Variables

We use a method called "permutation feature importance," which was introduced by Breiman (2001) to determine the importance of variables. The main idea behind this method is that if randomly permutating the value of a feature increases the model error, the feature is "important" because the model relies on this feature to make predictions. If the value of the feature is changed, and the model error remains the same, the feature is "not important." The algorithm is described in Fisher, Rudin, and Dominici (2018):

The inputs are: trained model f, feature matrix X, target vector y, error measure $L(y, f)$. The steps of the algorithm are:

1. Estimate the original model error $e^{orig} = L(y, f(X)$ (e.g., mean squared error).
2. For each feature $j = 1, \ldots, p$:
    - generate feature matrix X^{perm} by permuting feature j in the data X. This breaks the association between feature j and the true outcome y;
    - estimate error $e^{perm} = L(y, f(X^{perm}))$ based on the predictions of the permutated data;
    - calculate permutation feature importance $FI_j = e^{perm}/e^{orig}$.
3. Sort features by descending FI_j.

To illustrate how permutating a feature will increase the errors, we show the percentage of loss increase $(FI_j - 1 = (e^{perm} - e^{orig})/e^{orig})$ in Table 4. The order of importance of the variables is different for the three different transitions. Age plays the most important role in the health transitions of healthy individuals. Lifestyle behaviors are important for the onset of LTC needs. Gender is the most important for the mortality of the disabled, which is why many studies on mortality modeling often split the data by gender first. We do not need to do this prework because the neural network automatically detects these differences and relationships. Furthermore, marital status plays a relatively important role in all health transitions. We note that age and time are not very important in modeling the mortality of disabled individuals.

In summary, we find that several socioeconomic and health behavior factors influence the transition intensities significantly. However, few previous studies consider factors other than age and time when modeling health transition intensities. Estimates of the importance of different variables suggest that other than age and time, socioeconomic and lifestyle factors should also be considered in health transition estimations.

Table 4. Importance of Different Variables.

Variables	σ: H→L (%)	μ: H→D (%)	ν:L→D (%)
Age	72.98	200.63	1.77
Time	7.04	2.16	0.89
Gender	5.60	12.02	8.25
Residency	3.04	0.34	0.93
Marital status	7.78	7.00	4.26
Smoke	13.93	2.89	3.18
Drink	8.06	1.86	2.63

4.3. Proportional Hazards Model and Survival Curve

Our model can also estimate the hazard function in survival analysis. The hazard function is a measure of risk at time t and is typically denoted as $\lambda(t)$. If we assume that an individual has survived for a time t, and, then the hazard function is the probability that the individual will not survive for an additional time δ:

$$\lambda(t) = \lim_{\delta \to 0} P_r(t \leq T < t + \delta | T \geq t)/\delta \ . \tag{14}$$

Formula (14) is very similar to transition intensity (2), so our proposed model can be used to estimate the hazard function.

The proportional hazards model is a widely used method in survival analysis. The hazard function of the model consists of two functions: a baseline hazard function, $\lambda_0(t)$ and a risk function, $h(x)$, representing the influence of covariates x:

$$\lambda(t|x) = \lambda_0(t) * h(x), \tag{15}$$

$$\frac{\lambda(t|x)}{\lambda_0(t)} = h(x) = \exp^{\alpha x} , \tag{16}$$

where α stands for coefficients of x. Our proposed model can be used to estimate the hazard ratio $h(x)$ with age as the time scale.

We use the estimates of $h(x)$ from our proposed model to test if the proportional hazard assumption that the baseline hazard ratio $\lambda_0(t)$ is constant over the time scale (i.e., age in our case) is satisfied. Figure A.4 in the appendix tests shows that the proportional hazard assumption is violated for different variables.

Figure A.5 in the appendix shows the survival curve by time for all individuals aged 85 in 1998 in our data. We compare the percentage of individuals alive at all subsequent CLHLS survey waves for three models: GLM, CME, and CMET. Our proposed model CMET provides a good fit to the data.

4.4. Life Expectancy and Healthy Life Expectancy

One of the advantages of our new model CMET is that it can be used to calculate life expectancy for all demographic, socioeconomic, and lifestyle factors included in the model. To calculate life expectancy for specific values of specific variables, we use CMET to estimate and predict the health transition intensities for people with different characteristics. This allows

us to use the full sample containing three types of transitions. Then we take an average of life expectancy across all the other variables when we only consider one certain variable.

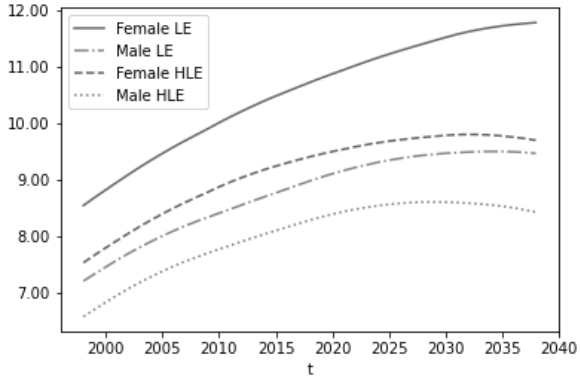
In the following, we provide estimates of life expectancy and healthy life expectancy for three different years: the year of the first CLHLS survey (1998), the year of the most recent CLHLS survey (2018), and 20 years after the last CLHLS survey (2038). Based on initial health status, we calculate life expectancy and healthy life expectancy at ages 75 and 85. We define an individual as "healthy" if the individual has no more than two ADL limitations, which is consistent with the definition in section 2.1. The resulting life expectancy for healthy and disabled individuals is shown in Table 5. Overall, life expectancy increases over time. For individuals who are healthy initially, the model predicts a nearly 1.5-year increase every 20 years for 75-year-old individuals and a one-year increase for 85-year-old individuals. The model predicts slower increases in life expectancy for the disabled. Healthy life expectancy increases over time as well. From 1998–2018, the healthy life expectancy for 75-year-old individuals and 85-year-old individuals increased by 1.8 and 1.3 years, respectively, and is projected to further increase over the next 20 years.

The life expectancy estimates also differ by other demographic, socioeconomic, and lifestyle factors. Figure 8 shows that women have a longer life expectancy than men. In terms of lifestyle behaviors, smokers and drinkers have a much lower life expectancy than non-smokers and non-drinkers. These differences grow over time. For socioeconomic factors, individuals who live in urban areas have a higher life expectancy, which is in line with other studies (e.g., Hanewald et al., 2019). Furthermore, healthy married individuals have a higher life expectancy than those who live alone. Healthy life expectancy shares a similar pattern to the life expectancy of healthy individuals by gender, smoking, drinking, residency, and marital status.

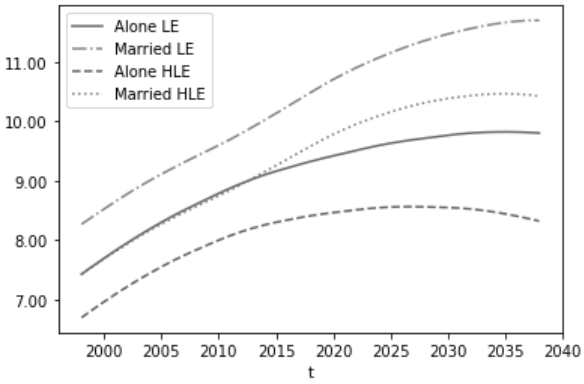
Table 5. Life Expectancy (LE) and Healthy Life Expectancy (HLE).

	75 years old			85 years old		
Variables	1998	2018	2038	1998	2018	2038
LE for healthy	8.27	9.77	11.25	5.10	6.26	7.14
HLE	7.08	8.85	10.57	4.37	5.66	6.62
LE for disabled	4.11	4.82	5.65	2.74	2.78	2.80

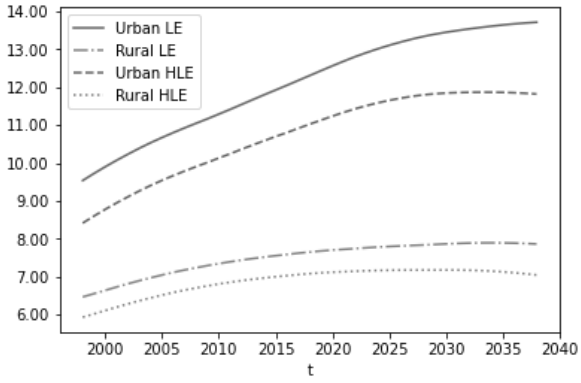
Figure 8. Life Expectancy (LE) and Healthy Life Expectancy (HLE) of Healthy Individuals.



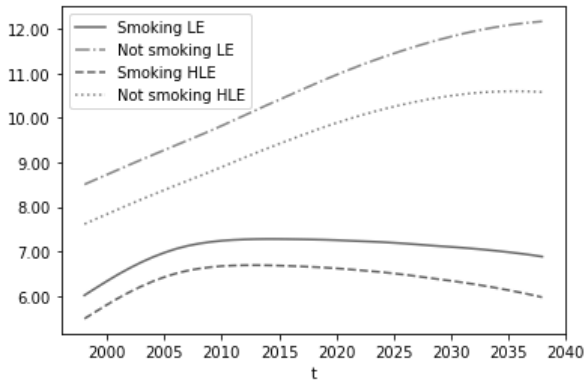
(a) LE and HLE by gender



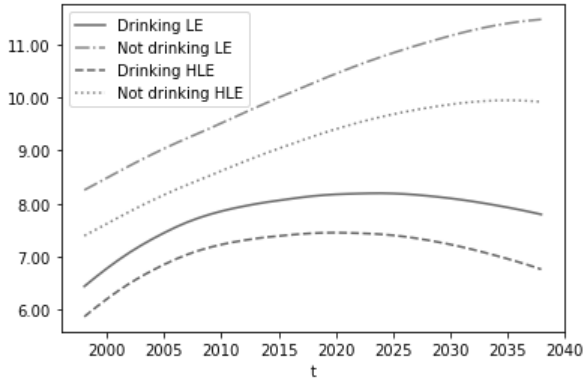
(b) LE and HLE by marital status



(c) LE and HLE by residency

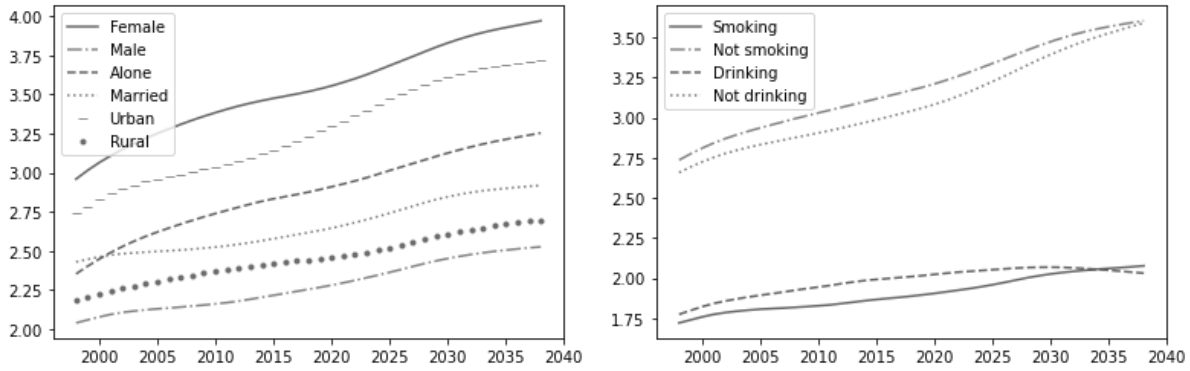


(d) LE and HLE by smoking



(e) LE and HLE by drinking

Figure 9. Life Expectancy (LE) of Disabled Individuals.



(a) LE by gender, marital status, and residency (b) LE by smoking and drinking

Using our new model CMET, we can calculate life expectancy for certain groups of individuals. In Table A.6 in the appendix, we report life expectancy for 75-year-old healthy individuals as an example. The table shows that individuals with different characteristics have different life expectancies and experience different increases in life expectancy over time.

Overall, our results show health differences among individuals with different backgrounds. We document improvements in life expectancy over time. The results confirm the adverse effects of smoking and drinking on health. Our model can be used to calculate and predict life expectancy for individuals with different characteristics.

5. Sensitivity Analysis

The results of the models will change if we split the dataset randomly. Therefore, the in-sample and out-of-sample performance of our model will be slightly different, depending on which data are used for training and testing. To evaluate the impact of the choice of data for the training set and test set on the performance of the models, we perform 100 trials using simple random sampling (SRS) to split the dataset. SRS is the most common method for data splitting and has been implemented in neural networks (May et al., 2010). Figure A.6 in the appendix shows the out-of-sample results of 100 trials for CMET.

We calculate the bias and variance of the testing error as follows:

$$E(\text{MSE}) = \frac{1}{M}\sum_{m=1}^{M} \text{MSE}, \tag{14}$$

$$V(\text{MSE}) = \frac{1}{M-1}\sum_{m=1}^{M}(\text{MSE}_m - E(\text{MSE}))^2 \quad (15)$$

Table A.7 in the appendix summarizes the bias and variance of the out-of-sample error for different models. We find that CMET has the lowest average MSE compared to other models. We also note that CMET has a relatively small variance of estimates for all intensities, which shows that our new model provides very stable estimates.

While the overall structure of the models influences performance, different choices of hyperparameters, including the number of layers, nodes, epochs, and batches, also affect the fitting and prediction process of the neural network. Table A.8 in the appendix shows a comparison of the different hyperparameters. We use as our default setting the settings of the CMET model analyzed in section 4.1: three hidden layers, 80 nodes, 100 epochs, and a batch size of 32. The results in Table A.8 are the averages of 100 times of training.

The results in Table A.8 show that a higher number of layers and nodes can improve the model's fit and reduce the in-sample losses. However, better fitting does not always mean better prediction performance: a model version with four hidden layers leads to higher out-of-sample losses for μ and ν than the base case with three hidden layers. Also, adding layers and nodes will increase complexity and is time-consuming. Larger epochs, which means increasing the number of iterations, or reducing the batch size, do not always improve the model but do increase the time. A higher out-of-sample loss for ν with a setting of 150 epochs compared to the default setting suggests that the estimate is not improved but requires more running time. Thus, the choice of hyperparameters of the neural network involves trade-offs between fitting and prediction, time, and performance.

6. Conclusions

This study proposes a new model that combines a neural network with a GLM to estimate and predict health transition intensities. Our model incorporates age effects, time trends, socioeconomic factors, and lifestyle behaviors in a Markov model with three health states (healthy, LTC needs, and dead). The model detects and incorporates complex relationships among the variables autonomously; that is, the model does not require these relationships to be specified in advance. We model different health transitions together using transfer learning, which also allows us to use the available data more effectively.

We illustrate the use of the new model based on data for individuals aged 65–105 from the

CLHLS over the period 1998–2018. We identify important factors explaining the transition intensities between the three health states for different subpopulations. We find that all of the variables mentioned above, including the socioeconomic and lifestyle factors, impact the onset of LTC needs and mortality of individuals with LTC needs. However, the ranking of the variables by importance varies for the different health transitions. The comparison of losses shows that basic combined GLM-NN models outperform most standalone GLM and neural network models. A combined model called CMET with expert opinion and transfer learning performs best.

We apply the new combined model to estimate life expectancy and healthy life expectancy and analyze how socioeconomic and lifestyle factors impact these measures. The results suggest that life expectancy will continue to increase over time, and that place of residence, marital status, and lifestyle factors such as smoking and drinking influence life expectancy significantly. Our model allows researchers and practitioners to predict the life expectancy and health expectancy of individuals with different backgrounds.

Overall, our study shows that the combination of traditional actuarial ideas and neural networks provides better-performing health transition models. Transfer learning enhances the performance of such combined models by linking the model estimation for different health transitions.

Our proposed new model has broad applications and provides a starting point for further health-related research. The new model can be easily applied to other health datasets. For example, the model can be used to estimate and predict health transitions with differently defined health states, a different number of health states, and for other countries. For instance, the model can be applied to fit and predict individual-level mortality data. We explained in section 2.1 that a version of the model with two health states (alive, dead) could be used to model mortality rates. While this would involve a single health transition (from alive to dead), transfer learning could be used to study the links between subgroups of individuals with different characteristics.

Our model can also be modified for classification problems where inputs are classified into categories of outputs. In this case, the activation function at the last layer should be replaced by a nonlinear activation function such as Softmax. With this change, the model could be used to predict an individual's health status directly. This change would also make it possible to apply our model to health insurance claims data for risk prediction. To adapt our model for claims data, the inputs of our model need to be replaced by information from medical claims,

including medical codes (diagnosis, procedure, medication) and individual-level information (age, sex, annual cost). Future research in this area could build previous research which has shown that deep learning in electronic health record data is promising (Lin et al., 2019).

Second, the model can be modified to improve micro-simulation health models, for example, the Future Elderly Model (FEM), which projects health and health care costs in the United States (Goldman et al., 2015) and the COMPAS model, which projects health trajectories in Canada (Boisclair et al., 2019). Our modeling strategies could be used to improve the individual-level health transition models in these models. Our model could allow researchers to incorporate additional variables and find linear and nonlinear relationships between the variables to improve the projection accuracy. Furthermore, our model with transfer learning could also be used to develop new micro-simulation health models for other countries, including countries with limited data, by transferring knowledge from developed models such as FEM and COMPAS.

Our model can also be extended to develop new multi-population mortality models, for example, by adding geographic variables (e.g., state or province). Transfer learning can discover the relationships between different subpopulations and improve model performance when data are limited in some areas. Transfer learning can help develop mortality models for populations with limited mortality data and address data limitations at older ages. Overall, transfer learning promises to be a useful tool for health and insurance studies when datasets are relatively small and when related datasets can be explored to offer meaningful information to the task under study.

Future research can also apply the model to pricing LTCI and other types of health insurance for individuals with different characteristics. It would also be interesting to use neural networks to determine the causality between socioeconomic variables and health transitions (see Chattopadhyay et al., 2019, for methods to detect causality with neural networks).

In summary, we believe that there are several promising directions for future research in insurance studies and actuarial science based on machine learning techniques.

References

Alashkar, T., S. Jiang, S. Wang, and Y. Fu, 2017, Examples-Rules Guided Deep Neural Network for Makeup Recommendation, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI Press: 941-947.

Alter, G., and J. C. Riley, 1989, Frailty, Sickness, and Death: Models of Morbidity and Mortality in Historical Populations, *Population Studies*, 43(1): 25-45.

Biessy, G., 2017. Continuous-Time Semi-Markov Inference of Biometric Laws Associated with a Long-Term Care Insurance Portfolio, *ASTIN Bulletin,* 47(2): 527-561.

Boisclair, D., Y. Décarie, F. Laliberté-Auger, & P. C. Michaud, 2019, COMPAS: A health microsimulation model for Quebec and Canada. Technical document, Chaire de recherche sur les enjeux économiques intergénérationnels, Montréal.

Breiman, L., 2001, Random Forests, *Machine Learning*, 45(1): 5-32.

Brockett, P. L., L. L. Golden, J. H. Jang, and C. H. Yang, 2006, A Comparison of Neural Network, Statistical Methods, and Variable Choice for Life Insurers' Financial Distress Prediction, *Journal of Risk and Insurance*, 73: 397-419.

Bu, Z., S. Xu, and K. Chen, 2021, A Dynamical View on Optimization Algorithms of Overparameterized Neural Networks. In *International Conference on Artificial Intelligence and Statistics* (pp. 3187-3195). PMLR.

Center for Healthy Aging and Development Studies, 2020, The Chinese Longitudinal Healthy Longevity Survey (CLHLS)-Longitudinal Data (1998-2018), https://doi.org/10.18170/DVN/WBO7LK, Peking University Open Research Data Platform, V2.

Chattopadhyay, A., P. Manupriya, A. Sarkar, and V. N. Balasubramanian, 2019, Neural Network Attributions: A Causal Perspective. In *International Conference on Machine Learning* (pp. 981-990). PMLR.

Cheng, H. T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, ... and R. Anil, 2016, Wide & Deep Learning for Recommender Systems, In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (pp. 7-10).

Cheng, X., Z. Jin, and H. Yang, 2020, Optimal Insurance Strategies: A Hybrid Deep Learning Markov Chain Approximation Approach, *ASTIN Bulletin*, 50(2): 449-477.

Christiansen, M. C., 2012, Multistate Models in Health Insurance, *AStA Advances in Statistical Analysis*, 96(2): 155-186.

Dai, A. M., and Q. V. Le, 2015, Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems* (pp. 3079-3087).

Eberhardt, M. S., and E. R. Pamuk, 2004, The Importance of Place Of Residence: Examining Health in Rural and Nonrural Areas, *American Journal of Public Health*, 94(10): 1682-1686.

Engchuan, W., A. C. Dimopoulos, S. Tyrovolas, F. F. Caballero, A. Sanchez-Niubo, H. Arndt, ... and D. B. Panagiotakos, 2019, Sociodemographic Indicators of Health Status Using a Machine Learning Approach and Data from the English lLongitudinal Study of Aging (ELSA), *Medical Science Monitor*, 25: 1994.

Fisher, A., C. Rudin, and F. Dominici, 2018, Model Class Reliance: Variable Importance Measures for Any Machine Learning Model Class, from the "Rashomon" Perspective, arXiv preprint arXiv:1801.01489, 68.

Fong, J. H., A. W. Shao, and M. Sherris, 2015, Multistate Actuarial Models of Functional Disability, *North American Actuarial Journal*, 19(1): 41-59.

Fuino, M., and J. Wagner, 2018, Long-Term Care Models and Dependence Probability Tables by Acuity Level: New Empirical Evidence from Switzerland, *Insurance: Mathematics and Economics*, 81: 51-70.

Gabrielli, A., 2020, A Neural Network Boosted Double Over Dispersed Poisson Claims Reserving Model, *ASTIN Bulletin*, 50(1): 25-60.

Gabrielli, A., Richman, R., & Wüthrich, M. V., 2020, Neural network embedding of the over-dispersed Poisson reserving model. *Scandinavian Actuarial Journal*, 2020(1), 1-29.

Goldman, N., S. Korenman, and R. Weinstein, 1995, Marital Status and Health Among the Elderly, *Social Science & Medicine*, 40(12): 1717-1730.

Goodfellow, I., D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, 2013, Maxout Networks,

In International Conference on Machine Learning (pp. 1319-1327), PMLR.
Gu D., 2008, General Data Quality Assessment of the CLHLS. In: Yi Z., D.L. Poston, D.A. Vlosky, D. Gu (eds) Healthy Longevity in China. *Demographic Methods and Population Analysis*, vol 20., Springer, Dordrecht.
Gu, D., and Q. Feng, 2015, Frailty Still Matters to Health and Survival in Centenarians: The Case of China, *BMC Geriatrics*, 15(1): 159.
Gu, D., and Z. Yi, 2004, Sociodemographic Effects on the Onset and Recovery of ADL Disability Among Chinese Oldest-Old, *Demographic Research*, 11: 1-42.
Hanewald, K., H. Li, and A. W. Shao, 2019, Modelling Multistate Health Transitions in China: A Generalised Linear Model with Time Trends, *Annals of Actuarial Science*, 13(1): 145-165.
Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, 2012, Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, arXiv preprint arXiv:1207.0580.
Husemoen, L. L. N., M. Osler, N. S. Godtfredsen, and E. Prescott, 2004, Smoking and Subsequent Risk of Early Retirement Due to Permanent Disability, *The European Journal of Public Health*, 14(1): 86-92.
Ikeda, T., T. Tsuboya, J, Aida, Y. Matsuyama, S. Koyama, K. Sugiyama, ... and K. Osaka, 2019, Income and Education Are Associated with Transitions in Health Status Among Community-Dwelling Older People in Japan: The JAGES Cohort Study, *Family Practice*, 36(6): 713-722.
Ioffe, S., and C. Szegedy, 2015, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv preprint arXiv:1502.03167.
Jeong, G., and H. Y. Kim, 2019, Improving Financial Trading Decisions Using Deep Q-learning: Predicting the Number of Shares, Action Strategies, and Transfer Learning, *Expert Systems with Applications*, 117: 125-138.
Johansson, S. R., 1991, The Health Transition: The Cultural Inflation of Morbidity During the Decline of Mortality, *Health Transition Review*, 39-68.
Jones, R. H., Xu, S., & Grunwald, G. K. (2006). Continuous time Markov models for binary longitudinal data. *Biometrical Journal*, *48*(3), 411-419.
Kiermayer, M., and C. Weiß, 2020, Grouping of Contracts in Insurance Using Neural Networks, *Scandinavian Actuarial Journal*, 1-28.
Kim, H., and B. D. Youn, 2019, A New Parameter Repurposing Method for Parameter Transfer with Small Dataset and Its Application in Fault Diagnosis of Rolling Element Bearings, *IEEE Access*, 7: 46917-46930.
Kingma, D. P., and J. Ba, 2014, Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980.
Klambauer, G., T. Unterthiner, A. Mayr, and S. Hochreiter, 2017, Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems* (pp. 971-980).
Kumagai, W., 2016, Learning Bound for Parameter Transfer Learning. In *Advances in Neural Information Processing Systems* (pp. 2721-2729).
May, R. J., Maier, H. R., & Dandy, G. C., 2010, Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks*, *23*(2), 283-294.
Li, C., S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, and A. Shilton, 2018, High Dimensional Bayesian Optimization Using Dropout, arXiv preprint arXiv:1802.05400.
Li, Z., A. W. Shao, and M. Sherris, 2017, The Impact of Systematic Trend and Uncertainty on Mortality and Disability in a Multistate Latent Factor Model for Transition Rates, *North American Actuarial Journal*, 21(4): 594-610.
Mu, H., and Z. Yang, 2016, A Study on the Effect of Marital Status on the Probability of Elderly Death: An Empirical Analysis Based on CLHLS Cohort Data, *Southern Population*, 04: 38-49 [in Chinese].
Nusselder, W. J., C. W. N. Looman, P. J. Marang-van De Mheen, H. Van de Mheen, and J. P. Mackenbach, 2000, Smoking and the Compression of Morbidity, *Journal of Epidemiology & Community Health*, 54(8): 566-574.

Oslin, D. W., 2000, Alcohol Use in Late Life: Disability and Comorbidity, *Journal of Geriatric Psychiatry and Neurology*, 13(3): 134-140.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever, 2018, Improving Language Understanding with Unsupervised Learning, Technical Report, OpenAI.

Renshaw, A. E., 1991, Actuarial Graduation Practice and Generalised Linear and Non-linear Models, *Journal of the Institute of Actuaries*, 118(2): 295-312.

Renshaw, A. E., and S. Haberman, 1995, On the Graduations Associated with a Multiple State Model for Permanent Health Insurance, *Insurance: Mathematics and Economics*, 17(1): 1-17.

Robards, J., M. Evandrou, J. Falkingham, and A. Vlachantoni, 2012, Marital Status, Health and Mortality, *Maturitas*, 73(4): 295-299.

Schelldorfer, J., and M. V. Wüthrich, 2019, Nesting Classical Actuarial Models into Neural Networks, available at SSRN 3320525.

Shao, A. W., M. Sherris, and J. H. Fong, 2017, Product Pricing and Solvency Capital Requirements for Long-Term Care Insurance, *Scandinavian Actuarial Journal*, 2017(2): 175-208.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *The Journal of Machine Learning Research*, 15(1): 1929-1958.

Tan, C., F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, 2018, A Survey on Deep Transfer Learning. In *International Conference on Artificial Neural Networks* (pp. 270-279). Springer, Cham.

Tang, S., Q. Liu, and W. A. Tan, 2019, Intention Classification Based on Transfer Learning: A Case Study on Insurance Data. In *International Conference on Human Centered Computing* (pp. 363-370). Springer, Cham.

United Nations, 2016. Briefing Paper: Growing Need for Long-Term Care: Assumptions and Realities. United Nations Department of Economic and Social Affairs Ageing. Retrieved from: https://www.un.org/esa/socdev/ageing/documents/un-ageing_briefing-paper_Long-term-care.pdf

Wei, M., and H. Wang, 2017, Gender, Urban-Rural and Cohort Differences in the Disability Trajectory of the Elderly in China, *Population & Development*, 23: 5 [in Chinese].

Wüthrich, M. V., and M. Merz, 2019, Yes, We CANN!, *ASTIN Bulletin*, 49(1): 1-3.

Ye, R., and Q. Dai, 2018, A Novel Transfer Learning Framework for Time Series Forecasting, *Knowledge-Based Systems*, 156: 74-99.

Yin, K. S., and S. S. Htay, 2020, Prediction of Natural Gas Final Consumption using Artificial Neural Networks. In *2020 International Conference on Advanced Information Technologies (ICAIT)* (pp. 224-229), IEEE.

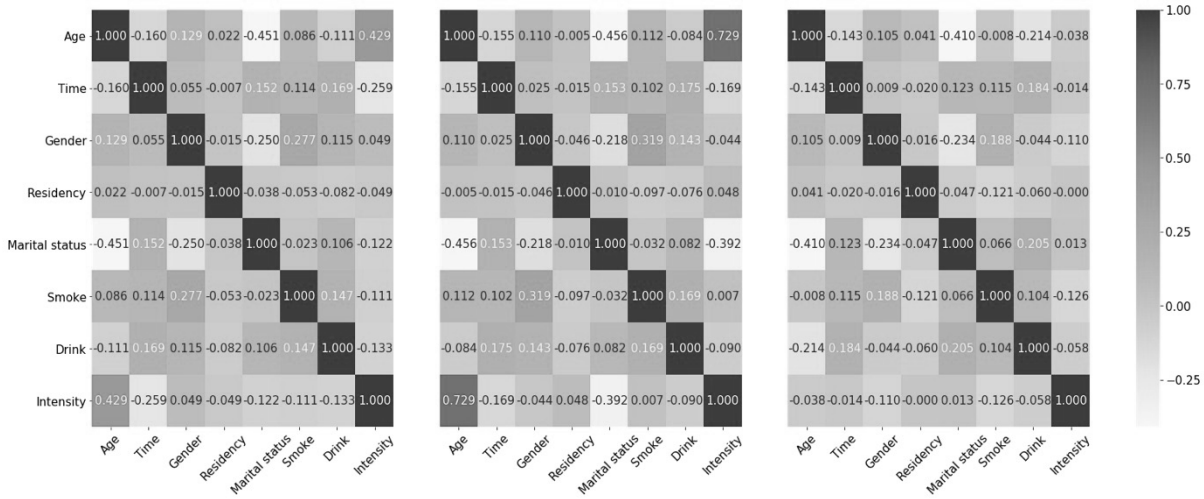
Zeng, Y., 2013, A Follow-Up Survey of Factors Affecting Health of the Elderly in China (1998–2012) and a Review of Related Policy Research (Part 2), *Aging Science Research*, (2): 63-71 [in Chinese].

Zheng, Z., 2020, Twenty Years' Follow-Up on Elder People's Health and Quality of Life, *China Population and Development Studies*, 1-13.

Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, 2020, A Comprehensive Survey on Transfer Learning, arXiv e-prints, arXiv:1911.02685.

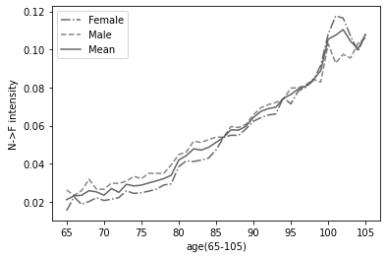
Appendix

Figure A.1. Correlations between Variables.

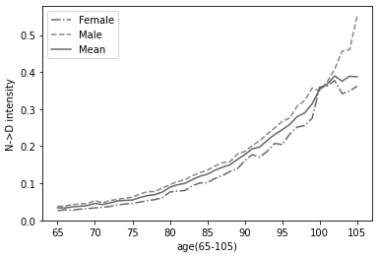


(a) σ: H→L (b) μ: H→L (c) ν: L→D

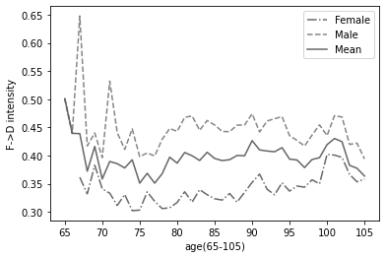
Figure A.2. Intensities by Ages and Different Variables.



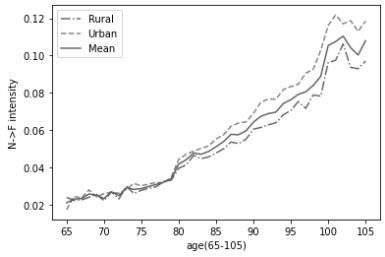
(a) σ: N→F by gender



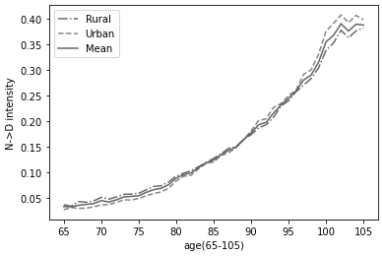
(b) μ: N→D by gender



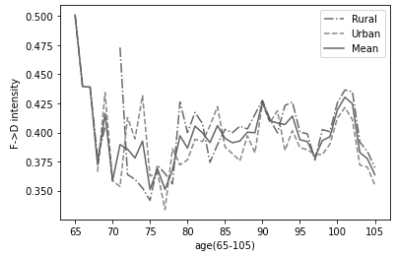
(c) ν: F→D by gender



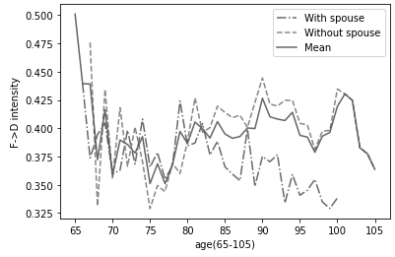
(d) σ: N→F by residency



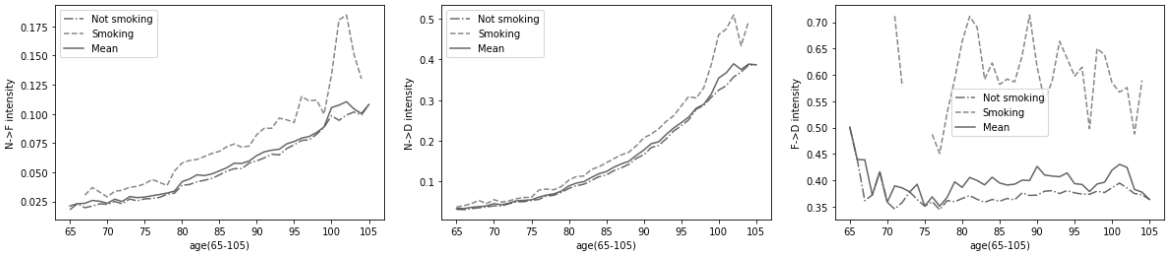
(e) μ: N→D by residency



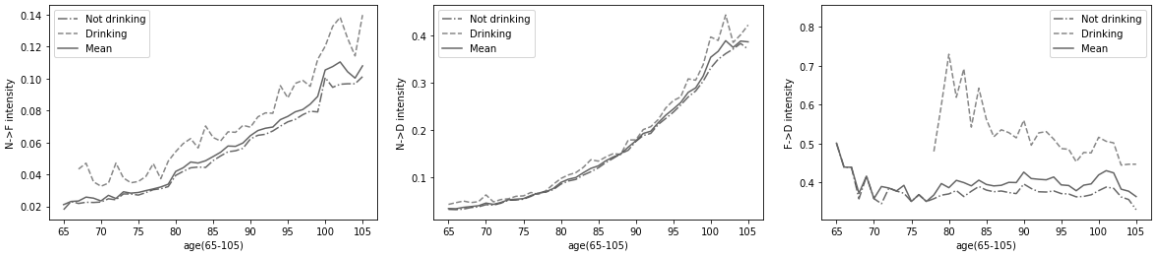
(f) ν: F→D by residency



(g) σ: N→F by marital status (h) μ: N→D by marital status (i) ν: F→D by marital status

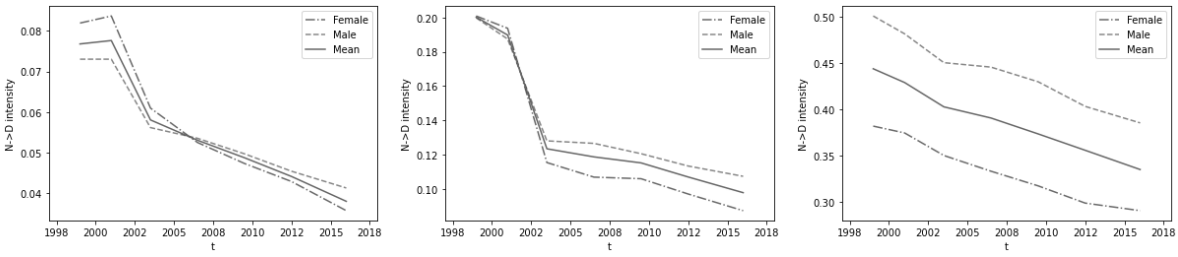


(j) σ: N→F by smoking (k) μ: N→D by smoking (l) ν: F→D by smoking

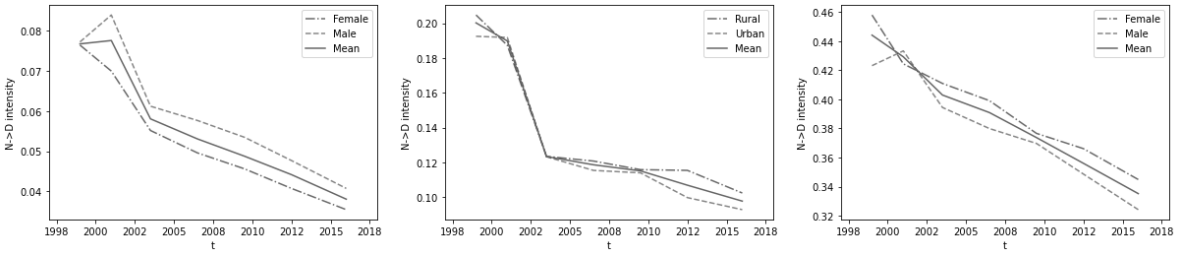


(m) σ: N→F by drinking (n) μ: N→D by drinking (o) ν: F→D by drinking

Figure A.3. Intensities with Time for Different Variables.



(a) σ: N→F by gender (b) μ: N→D by gender (c) ν: F→D by gender



(d) σ: N→F by residency (e) μ: N→D by residency (f) ν: F→D by residency

Figure A.4. –ln(–ln) Survival By Different Variables.

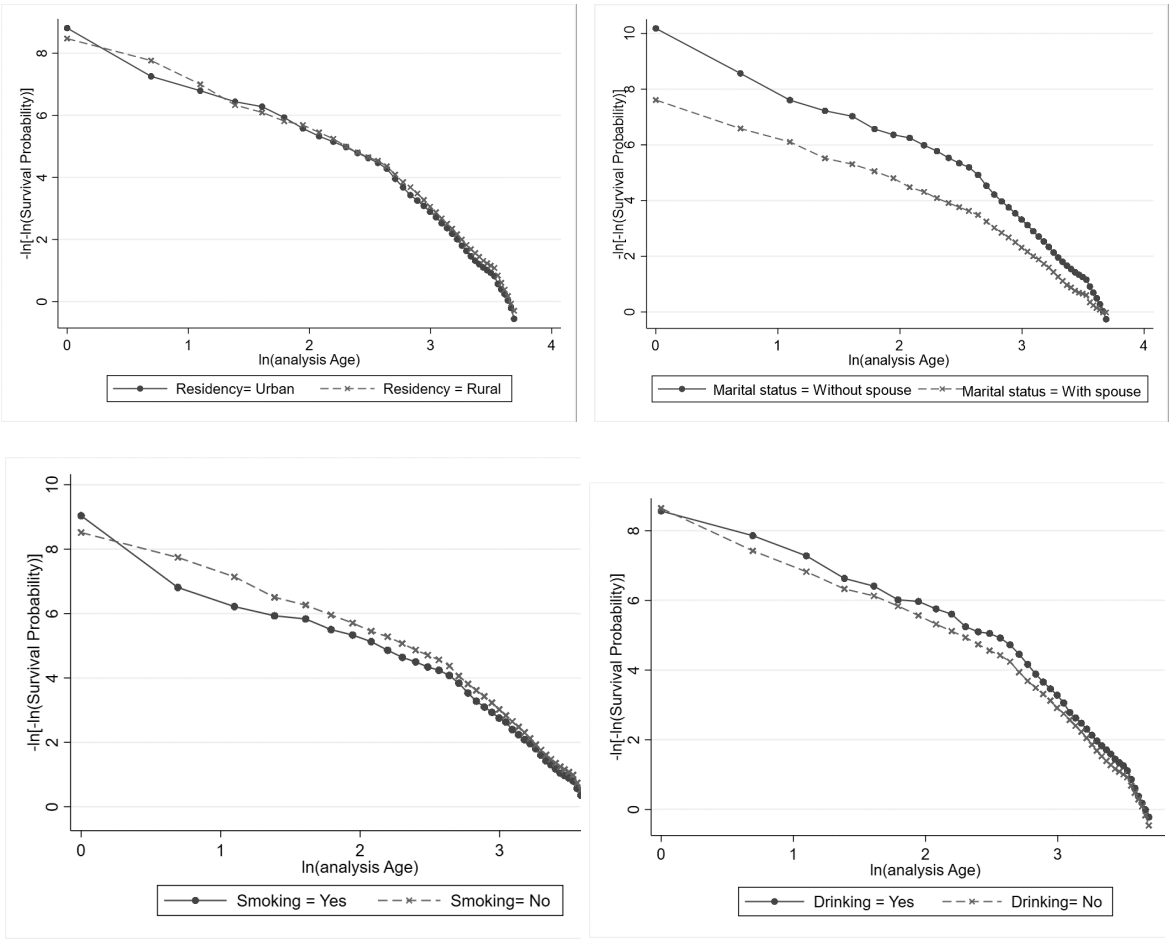


Figure A.5. Survival Curve for Individuals Aged 85 in 1998.

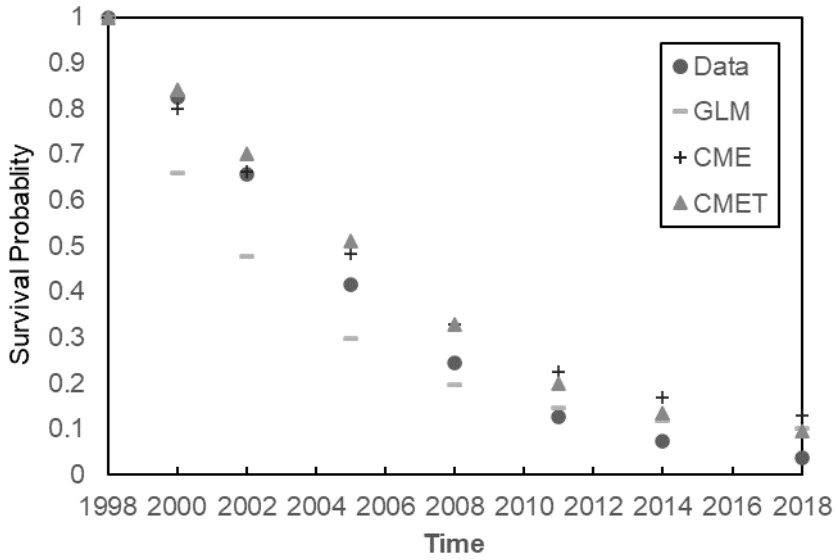


Figure A.6. Out-of-Sample Losses (× 10-^{2}) of 100 Trials of Different Transitions.

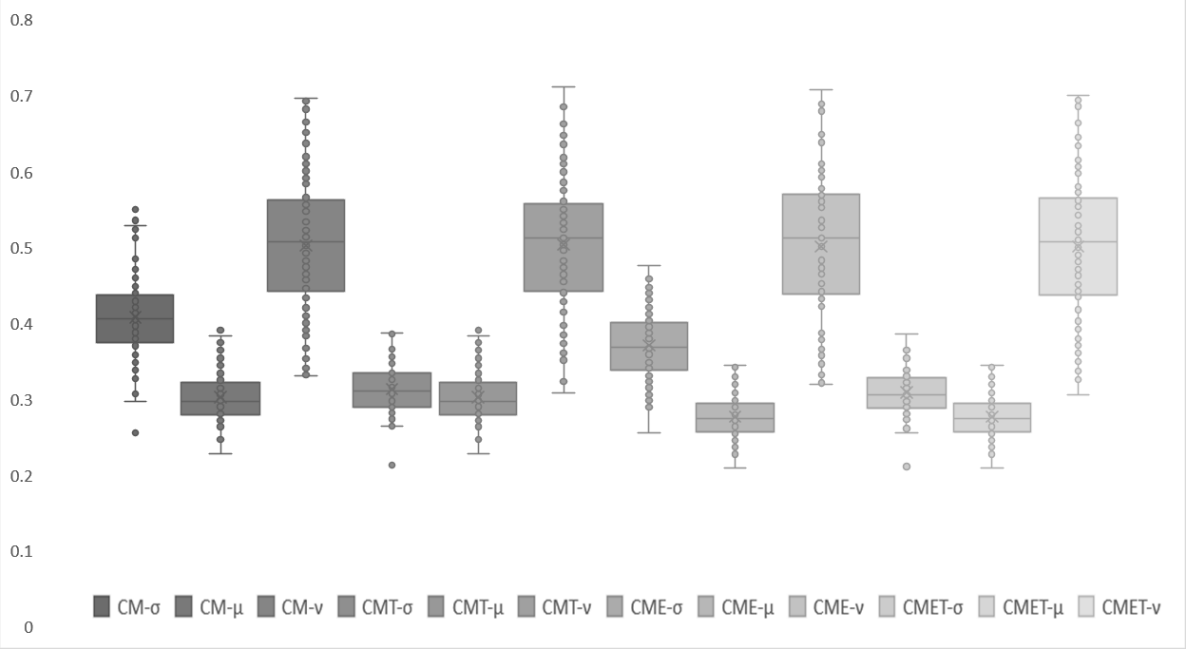


Table A.1. Attrition Rates.

Wave	Full sample		
	Individuals	Lost to follow-up compared to the previous wave	Attrition rate
2000	9,093	894	9.83%
2002	11,199	1,541	13.76%
2005	16,064	2,015	12.54%
2008	15,638	2,938	18.79%
2011	16,954	2,894	17.07%
2014	9,765	820	8.40%
2018	7,192	1,525	21.20%

Note: Numbers obtained from the CLHLS website. Numbers refer to the full CLHLS data.

Table A.2. Transition Counts for Different Variables at Ages 85 and 105.

		Transition counts (age 85/105)						Exposure years (age 85/105)			
		σ: H→L		μ: H→D		ν: L→D		H		L	
Age		85	105	85	105	85	105	85	105	85	105
Gender	Male	74	8	350	55	72	37	2,526	143	232	67
	Female	101	43	231	222	77	192	2,337	556	375	451

Marital status	With spouse	55	0	165	7	44	4	1,521	25	192	12
	Without spouse	120	51	416	270	105	225	3,341	674	415	506
Residency	Rural	93	24	339	181	63	128	2,727	466	277	286
	Urban	82	27	242	96	86	101	2,136	232	330	231
Smoke	Yes	29	4	139	31	16	10	998	91	84	22
	No	146	47	442	246	133	219	3,865	608	523	495
Drink	Yes	25	7	122	53	9	32	1,046	151	74	66
	No	150	44	459	224	140	197	3,817	547	533	451
Total		175	51	581	277	149	229	4,863	698	607	518

Table A.3. Crude Intensities for Time and Age Groups.

	Age / Time	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100–105
σ: H→L	1998–2000	-	-	4.15	3.71	6.29	7.45	9.05	9.57
	2000–2002	-	-	9.33	4.36	6.94	9.40	12.23	14.49
	2002–2005	0.50	0.92	1.36	2.41	3.58	5.13	5.80	6.95
	2005–2008	0.53	0.64	1.07	2.05	3.02	4.02	4.35	4.86
	2008–2011	0.73	1.16	1.71	2.88	4.25	6.01	6.82	6.87
	2011–2014	0.88	0.73	1.53	3.05	4.05	5.42	6.11	8.39
	2014–2018	0.35	0.73	1.64	1.75	3.01	4.60	4.96	4.54
μ: H→D	1998–2000	-	-	5.93	9.05	13.15	22.44	27.29	33.70
	2000–2002	-	-	9.33	10.29	14.59	19.29	31.82	37.51
	2002–2005	2.77	3.47	6.27	10.87	15.91	22.72	28.95	35.45
	2005–2008	2.30	4.11	4.87	9.01	15.27	21.74	27.04	34.95
	2008–2011	1.62	3.49	6.01	9.02	15.18	22.79	30.49	37.17
	2011–2014	1.52	2.46	4.58	7.63	12.49	19.76	26.39	24.98
	2014–2018	1.93	2.99	4.20	7.09	11.31	19.26	26.35	28.73
ν: L→D	1998–2000	-	-	21.18	16.20	18.86	28.45	44.09	48.46
	2000–2002	-	-	48.32	23.86	25.87	27.80	38.79	51.90
	2002–2005	11.72	9.94	11.03	19.41	26.72	32.56	39.39	43.81
	2005–2008	6.48	14.79	15.51	18.06	24.99	26.98	36.26	45.55

	2008–2011	6.00	13.35	12.31	12.88	18.87	29.09	37.31	41.81
	2011–2014	0.00	17.55	12.28	16.66	23.12	35.04	40.06	36.14
	2014–2018	14.41	18.80	14.58	18.64	22.89	25.33	40.27	46.98

Note: The numbers are all percentages.

Table A.4. GLM0 Regression Results.

	Coef for σ	Coef for μ	Coef for v
Age	-4.0799**	-2.1669***	-1.1757***
Time	-9.4054***	-6.7797***	-3.8519***
Age × Time	12.5894***	9.2766***	4.6225***

Note: *p<0.05; **p<0.01; ***p<0.001

Table A.5. GLM Regression Results.

	Coef for σ	Coef for μ	Coef for v
Age	-0.1519**	0.5834***	0.1728***
Time	-1.0198***	-0.7456***	-0.2935***
Gender_Male	-0.7331***	-0.4382***	0.0695**
Residency_Rural	-0.4864***	-0.3093***	-0.0752***
Marital status_With spouse	-0.3584***	-0.8153***	-0.1681***
Smoking_No	-1.1695***	-0.8979***	-0.6943***
Drinking_No	-0.8196***	-0.5876***	-0.2983***

Table A.6. Life Expectancy of Healthy 75-Year-Old Individuals.

			Male		Female	
75			Rural	Urban	Rural	Urban
1998	Smoke	Yes	7.03	7.45	7.32	7.27
		No	7.59	8.02	9.54	9.47
	Drink	Yes	7.31	7.44	8.33	7.96
		No	7.38	8.01	9.48	9.40
	Marital	Yes	7.57	8.26	9.58	8.37

		No	7.14	7.35	9.33	8.36
2018	Smoke	Yes	7.68	8.45	9.02	9.02
		No	9.55	10.20	11.06	11.69
	Drink	Yes	8.28	8.62	9.96	9.34
		No	8.94	10.06	11.00	11.66
	Marital	Yes	9.28	10.56	12.58	13.63
		No	8.14	8.64	10.54	11.07
2038	Smoke	Yes	9.09	9.69	9.68	9.55
		No	11.44	11.36	11.86	12.54
	Drink	Yes	9.99	9.99	10.87	10.58
		No	10.61	11.16	11.78	12.43
	Marital	Yes	10.98	11.67	14.54	15.25
		No	9.90	9.89	11.13	11.68

Table A.7. Bias and Variance Estimations for Different Models

Methods	E ($\times 10^{-2}$)			V ($\times 10^{-4}$)		
	σ: N→F	μ: N→D	ν: F→D	σ: N→F	μ: N→D	ν: F→D
GLM	81.09	86.05	55.09	60.72	42.24	92.23
NN	40.68	30.24	50.42	30.74	9.17	87.56
CM	40.83	30.26	50.24	29.50	11.20	82.17
CMT	31.38	30.26	50.38	8.82	11.20	85.79
CME	37.14	27.75	50.12	19.52	8.67	91.18
CMET	30.90	27.75	50.11	8.67	8.67	88.48

Table A.8. Hyperparameters Comparison.

	In-sample loss ($\times 10^{-2}$)			Out-of-sample loss ($\times 10^{-2}$)			Time (s)		
	σ: N→F	μ: N→D	ν: F→D	σ: N→F	μ: N→D	ν: F→D	σ: N→F	μ: N→D	ν: F→D

Hidden layers	2	36.23	36.55	54.16	33.61	33.03	59.11	9.76	14.58	8.89
	3	30.76	27.51	50.92	30.90	27.75	50.11	11.73	17.06	10.82
	4	25.55	25.43	50.75	30.45	29.50	52.10	12.50	18.52	11.44
Nodes	30	39.15	50.34	52.96	41.60	49.83	49.53	10.89	15.52	10.10
	80	30.76	27.51	50.92	30.90	27.75	50.11	11.73	17.06	10.82
	130	29.52	25.34	50.35	29.30	25.31	50.93	12.85	19.12	11.57
Epochs	50	33.42	34.96	53.98	32.62	38.32	51.92	8.59	11.22	8.10
	100	30.76	27.51	50.92	30.90	27.75	50.11	11.73	17.06	10.82
	150	29.95	25.75	50.84	29.26	25.82	51.72	14.53	21.95	13.64
Batch size	16	29.89	25.84	50.88	28.86	25.86	54.46	25.09	36.15	23.27
	32	30.76	27.51	50.92	30.90	27.75	50.11	11.73	17.06	10.82
	64	38.14	49.40	53.39	37.41	49.43	53.22	5.29	7.94	4.87